CENTER FOR GLOBAL DEVELOPMENT

Property Tax Compliance in Tanzania: Can Nudges Help?

Matthew Collin, David K. Evans, Vincenzo Di Maro, Fredrick Manang

Abstract

Low tax compliance in low- and middle-income countries around the world limits the ability of governments to offer effective public services. This paper reports the results of a text message campaign aimed at promoting tax compliance among landowners in Dar es Salaam, Tanzania. Landowners were randomly assigned to one of four groups designed to test different aspects of tax morale. They received a simple text message reminder to pay their tax (a test of salience), a message highlighting the connection between taxes and public services (reciprocity), a message communicating that people who did not pay were not contributing to local or national development (social pressure), or no message (control). Recipients of any message were 18 percent (or 2 percentage points) more likely to pay any property tax by the end of the study period. While each type of message resulted in gains in payment rates, total payment amounts were highest for those who received reciprocity messages. The average estimated benefit-cost ratio across treatments is 36:1 due to the low cost of the intervention, with higher cost-effectiveness for reciprocity messages. However, there was significant geographic heterogeneity in both treatment effect sizes and estimated cost-effectiveness across the city, highlighting potential gains to program effectiveness in cases where such heterogeneity can be predicted ex-ante.

KEYWORDS
Tax Compliance, Tax Morale, Public Finance, Nudges

JEL CODES
H26, H13, O17

This working paper was first published in August 2022. The original version is available here.

WORKING PAPER 621 · JULY 2024

Property Tax Compliance in Tanzania: Can Nudges Help?

Matthew Collin
EU Tax Observatory

David K. Evans
Inter-American Development Bank; formerly with the Center for Global Development

Vincenzo Di Maro
World Bank

Fredrick Manang
University of Dodoma

The authors are grateful for support from the Tanzania Revenue Authority (TRA), especially the non- tax revenue department at TRA headquarters. We thank Anne Brockmeyer, Santiago de la Cadena Becerra, Addisu Lashitew, Fabrizio Santoro, Stuti Khemani, Thiago Scot, and participants at the CSAE conference and at World Bank workshops for discussions and comments. We thank the DIME Analytics team at the World Bank for their careful replication review. The Impact Evaluation to Development Impact (i2i) multi-donor trust fund at the World Bank provided funding. The authors declare no competing interests. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not reflect the views of the Executive Directors of the World Bank or the governments they represent, nor of the employers of non-World Bank authors.
Corresponding author: Collin (mattcollin@gmail.com). Other authors: Manang (fredrick.manang@udom.ac.tz), Di Maro (vdimaro@worldbank.org), and Evans (devans@iadb.org).

Matthew Collin, David K. Evans, Vincenzo Di Maro, Fredrick Manang. 2024. "Property Tax Compliance in Tanzania: Can Nudges Help?" CGD Working Paper 621. Washington, DC: Center for Global Development. https://www.cgdev. org/publication/property-tax-compliance-tanzania-can-nudges-help

CENTER FOR GLOBAL DEVELOPMENT
2055 L Street, NW Fifth Floor
Washington, DC 20036
202.416.4000

1 Abbey Gardens
Great College Street
London
SW1P 3SE

www.cgdev.org

1 Introduction

Governments need revenue to fund public goods and services. The source of that revenue also matters (Gadenne 2017): the state's ability to raise tax revenue is thought to invite public scrutiny and strengthen the "social contract" between citizen and state, leading to positive effects on institutional and economic development (Besley and Persson 2013; Ali, Fjeldstad, and Katera 2017; Dincecco and Katz 2014). Despite the purported benefits of higher tax capacity, low- and middle-income countries largely struggle to raise the same levels of tax revenue as their higher income peers. As of 2014, the tax-to-GDP ratio of the median low or lower-middle income country was 15.8 percent, compared to 20.5 percent for higher income countries. Evidence suggests that this may condemn some of these countries to lower levels of economic growth in the long term (Gaspar, Jaramillo, and Wingender 2016).

One area where governments seek to improve tax compliance is the taxation of immovable property, which presents an attractive source of revenue as it is—in theory—both easier to target and easier to tax in a non-distortionary way. But low- and middle-income countries perform even worse in the collection of property tax than they do overall: where richer countries bring in around two percent of GDP in revenue from property taxes, lower income countries bring in less than one percent (Ali, Fjeldstad, and Katera 2017).

Tanzania, the context of our study, has historically struggled with low rates of tax revenue, both overall (11.8 percent of GDP) and for property tax (0.1 percent of GDP). Furthermore, the government has oscillated between a regime of decentralized revenue collection (where the local authorities are responsible for collecting property tax) and one of centralized collection (where the Tanzania Revenue Authority is responsible) (Fjeldstad, Ali, and Katera 2017). This process has led to unstable and unpredictable levels of property tax revenues over the past decade. Furthermore, the Tanzanian government has struggled with low levels of compliance. This may in part be due to a lack of perceived reciprocity by taxpayers: property owners do not understand how the government will use their money (PO-RALG 2013).[1] It is also the result of a small tax base: while legally every property owner is obligated to pay tax, local authorities have previously prioritized those with larger properties living in the most affluent areas of the city (PO-RALG 2013). It is within this context that the Tanzania Revenue Authority (TRA) is examining new ways to improve property tax compliance in the cities for which it is responsible.[2]

This paper investigates the impact of a series of reminders via text message that leverage different aspects of citizens' voluntary motivation to pay property taxes (i.e., their "tax morale"). Working with the TRA, we randomly allocated more than 200,000 individuals in Dar es Salaam who were liable to pay property tax—but as of one month prior to the annual deadline had not paid any—into four groups.[3] Three groups received a text message treatment: a simple reminder (increasing the salience of tax paying), a message that emphasized the link between tax revenue and publicly provided goods (focusing on reciprocity), or a message highlighting the

[1] In Pakistan, even an explicit intervention to collect citizen preferences on public services and then to deliver services based on those preferences had little initial effect on tax payments (Khwaja et al. 2020).

[2] Other studies have examined the overall effectiveness of tax units within government, such as Santoro and Waiswa (2023) in Uganda.

[3] As discussed later, taxpayers were actually assigned to five groups, but two groups received the same treatment due to an implementation error.

non-cooperative nature of tax evasion (focusing on social pressure). The fourth group served as a control.

We find that all three treatments had a positive impact on both the propensity of taxpayers to make payments to the TRA and the total amount paid. Those receiving the reciprocity treatment paid higher amounts of total tax. Similar to what has been found with some other "nudge" style interventions, the intervention is highly cost effective, with the increase in revenue driven by treatment exceeding its cost by roughly a factor of 36.

Finally, we document two interesting sources of heterogeneity. The first is across the geography of the city: areas that have a higher rate of tax compliance among the control group also had lower treatment effects, suggesting that nudges may be more successful in low-compliance areas when policymakers are able to identify them ex ante. Second, we find heterogeneity across the amount that taxpayers owed the TRA: average treatment effects appear to be strongest for those who owe less than 25,000 TSh, predominantly smaller properties that do not bring in much revenue and so are less likely to be subject to TRA follow up. This indicates that tax nudges may, in some circumstances, be regressive in their impact on compliance.

The rest of the paper is structured as follows. Section 2 discusses prevailing theories of tax compliance and the existing body of evidence on the impact of messages on taxpayer behavior. Section 3 discusses the structure of the experiment and the data we will use to examine its impact. Section 4 presents and discusses the results, and we conclude with Section 5.

2 Background

Until recently, economists have viewed tax compliance largely through the lens of enforcement (Allingham and Sandmo 1972), where taxpayers increase their compliance when the perceived probability of detection goes up. There is evidence that letters and electronic forms of communication have the potential to do this: research in many high-income economies suggests that letters containing an implicit or explicit threat of audit increases tax payments (Coleman 1996; Blumenthal et al. 2001; Hasseldine et al. 2007; Kleven et al. 2011; Fellner et al. 2013; Castro and Scartascini 2015; Hallsworth et al. 2017; Pomeranz 2015; Meiselman 2018; Hernandez et al. 2017), although the effect sizes vary across contexts and are not always significant (Ariel 2012).

Research in low- and middle-income countries has largely revealed similar results (Ortega and Scartascini 2020; Brockmeyer et al. 2019; Kettle et al. 2016; Shimeles et al. 2017; Brockmeyer et al. 2020; Holz et al. 2023), again with results not always significant (Del Carmen, Espinal Hernandez, and Scot 2022). Evidence from Rwanda suggests that text messages, the medium used in our study, are more effective than traditional paper letters in a lower income setting. That same study finds that less aggressive messages (such as reciprocity or reminder-framed messages) work better than those aimed at deterrence (Mascagni and Nell 2022), whereas evidence from Uganda suggests stronger impacts for enforcement focused messages (Cohen 2020). Thus, the effectiveness of different messages may depend on local norms. In a study in Tanzania (albeit in a different part of the country than our study), both direct threats of fines sent via text message and indirect threats passed on through local leaders led to an increase in payments by Tanzanian property owners (Mwaijande et al. 2021).

In recent years, economists have extended the Allingham and Sandmo model to include the concept of "tax morale," a bundle of mechanisms which explain voluntary tax compliance (Luttmer and Singhal 2014). Recent experiments have attempted to make the components of tax morale more salient through careful messaging, with mixed results. Kettle et al. (2016) find that both letters emphasizing national pride and those emphasizing social norms improve compliance in Guatemala, but not significantly more than a letter invoking a heightened probability of audit. In Peru, letters emphasizing social norms increased property tax compliance by twenty percent, twice as much as simple reminders (Del Carpio 2022). In Papua New Guinea, Hoy, Sinning, and McKenzie (2022) find that messages aimed at emphasizing the social benefits associated with taxes perform no better than reminders aimed at lowering the cost of compliance. In richer countries, randomized studies of letter or e-mail campaigns typically find that attempts to emphasize the social contract or civic duty either have little impact or are marginally effective (Coleman 1996; Blumenthal et al. 2001; Torgler 2004; Ariel 2012; Fellner et al. 2013; Castro and Scartascini 2015; Meiselman 2018) with some exceptions (Hallsworth et al. 2017). Krause (2020) finds that messages that emphasize the social pressure mechanism in Haiti might even have a negative effect on tax compliance. This is consistent with results from psychology research which have shown that in contexts with low rates of pro-social coordination, a mechanism known as antisocial punishment could be at play (Herrmann et al. 2008).

A recent meta-analysis (Antinyan and Asatryan 2020) of studies of nudges for tax compliance finds that on average deterrence nudges (those focused on audits or penalties) are effective at boosting compliance, whereas tax morale nudges are not. The meta-analysis also finds that nudges seem to work better for the sub-samples of late payers, which is the sample we focus on in this paper. Okunogbe and Santoro (2023) put electronic nudges within the context of broader efforts to incorporate technology into tax collection across African countries.[4]

3 Experiment and data collection

3.1 Baseline data and randomization

The frame for this experiment is a list of 241,200 properties for which, as of June 1, 2018, no property tax had been paid to the TRA for the 2017/2018 tax year. The deadline for property tax payments to be completed was June 30th. After June 30 had passed, the TRA extended the deadline for another two weeks, although it continued to accept payments after this point.

As some taxpayers own multiple parcels, we collapsed these data to the taxpayer level (237,699 taxpayers), using the taxpayer identification number associated with the property. We use two sources of information in the randomization: the location of the property and whether or not the property had been served a 'demand notice' at the time the data was collected. The location of the property is the lowest level of administration in Dar es Salaam, the sub-ward or 'mtaa' level. We assign taxpayers the same location as their property. When taxpayers have multiple properties that span more than one sub-ward, we pick the modal sub-ward. Where there is no modal sub-ward, we randomly assign one of the sub-wards to the taxpayer.

[4]Other work has explored efforts to strengthen property tax collection in other ways, such as how to match tax collectors with households to optimize payment (Bergeron et al. 2022).

Ultimately, the randomization was conducted within 1,211 different strata, which were defined both by the location of the property (sub-ward) and whether a bill had been issued.

Demand notices are bills issued by the TRA to landowners. Approximately 19 percent of the experimental sample had been issued a bill at the time of the data collection. The TRA issued bills (called "demand notices") in bulk for a specific area of the city (this could be a ward or a set of core streets) and sent them to landowners via manual delivery by TRA officers and temporary interns, typically after seeking the support of street leaders. While the goal is to cover all areas of the city every fiscal year, limited financial and human resources explain why only a subset of the city is covered in practice.

To better understand how our experimental sample compared to the average property in Dar es Salaam, we matched our experimental data to a set of data comprising every parcel in TRA's database for the city. Out of approximately 830,000 parcels in the city, nearly 30 percent are owned by a taxpayer in our experimental sample. Parcels included in our experiment were only slightly (1.4 percentage points) less likely to have been issued a bill. Of those that were issued bills, only 35 percent of properties in the experimental sample had been valued by the TRA, as opposed to 74 percent in the rest of the city (those that were not valued were charged a TSh 10,000 flat tax). Conditional on being billed and rated, the median value for properties included in the experiment was higher, approximately 31.2 million TSh (13,500 USD) versus 20 million TSh (8,600 USD) for the rest of the city. Properties that were billed and rated faced, on average, an annual tax rate of approximately 0.16% of the property's rated value.

3.2 Treatments

In collaboration with the Tanzania Revenue Authority, we randomized each property owner into one of five groups.[5] The treatments are summarized in Table 1. The control was not to receive any message from the TRA. Group T1 received a simple message reminding them to pay their property tax, indicating the due date (June 30th) and providing information the taxpayers could use to contact the TRA in case they had any questions. All other treatments included this reminder message.

Treatment 2 added a 'reciprocity' message, where taxpayers were reminded that taxes fund social services and infrastructure and finished with the TRA's slogan "Together we build our Nation." Treatment 3 included the simple reminder as well as a 'social pressure' message in which taxpayers were reminded, in a negative fashion, that non-compliers were not contributing to the development of the country or their own communities. The final planned treatment was an enforcement message which included the simple statement "Pay your tax early to avoid penalties." However, for reasons we discuss below, during the implementation of the experiment no taxpayers were sent this message.

[5]The randomization was conducted using the Stata command `randtreat`, with "misfits" (i.e., observations beyond those that are a multiple of the number of treatment groups) being dealt with using the *strata* method, which randomly allocates misfits across all strata (Carril et al. 2017).

Table 1: **Treatment arms and treatment assignments**

Treatment	Type	Message	N
Control	No message		47,555
T1	Reminder	"Dear taxpayer, TRA reminds you to pay your property tax. Pay before 30th June. For more information: dial * 152 * 00 #, visit your nearest TRA office or call 0800780078. Thank you."	47,502
T2	Reciprocity	[Reminder] + "Your tax facilitates access to social services and infrastructure. Together we build our nation."	47,547
T3	Social Pressure	[Reminder] + "Non-taxpayers are not contributing to national development and thus hindering development of their communities."	47,538
T4	Enforcement*	[Reminder] + "Pay your tax early to avoid penalties."	47,553

***Note:** Subjects randomly allocated to T4 were mistakenly sent T1. See sub-section 3.3 for details. Original messages were sent in Swahili.

3.3 Implementation and challenges

Following the randomization, a list of taxpayers and their phone numbers (included in the original TRA dataset) were provided to the TRA. The majority of messages were sent out after June 20, fewer than ten days before the initial deadline to pay property tax. While there was overlap in the delivery of different treatments, completion of each treatment arm proceeded sequentially, with reminder messages being sent first to group T1, reciprocity messages second to group T2, followed by social pressure messages to group T3.[6]

There were two errors in the message delivery process. First, the firm in charge of sending the messages sent Treatment 1 messages to taxpayers who had been randomly allocate to Treatment 4 (Enforcement), essentially doubling the size of the first treatment arm. Thus no enforcement messages were sent as a part of the experiment (although a few were sent independently by the TRA as part of a limited, non-experimental information campaign that happened concurrently). Second, instead of using the list of cleaned and prepared phone numbers they were provided, the firm chose an unformatted list which contained the same numbers, but in some cases were not usable due to missing pre-fixes or mistakenly included county codes. As a result, between 22-33 percent of each treatment arm was not sent the intended message. Using data from the text message delivery, we can account for which taxpayers were or were not sent a message due to this error.

Finally, the randomization was conducted at the taxpayer identification level, but a small subset of taxpayer IDs shared identical phone numbers. This is likely because some taxpayers were issued multiple taxpayer IDs, or households sharing a single number contained multiple taxpayers. This leads to spillovers in actual treatment between the various treatment arms and the control group.

These spillovers complicate our ability to discern the effects of each treatment in isolation. To account for this, for our main analyses we restrict our sample to the roughly 60% of taxpayers that have a unique phone number (one that is not shared with any other taxpayer in the sample). As we show in Appendix section A1.2, those who only have unique phone numbers are no more

[6] Figure A1 in the Appendix shows the timing of message delivery for the experimental sample. One concern is that the differential timing of message delivery affected each treatment's impact. In the results section, we provide suggestive evidence that our pattern of results is largely robust to this.

Table 2: Frequencies of actual treatment, by treatment group

Treatment Arm	Received No Message	Received T1	Received T2	Received T3
Control	99.99%	0%	0%	0%
T1 - Reminder	11.1%	88.9%	0%	0%
T2 - Reciprocity	4.3%	0%	93.01%	2.7%
T3 - Social pressure	15.18%	.01%	2.7%	82.13%

Note: As some some taxpayers received multiple messages, the above frequencies can exceed 100. Sample restricted to taxpayers with unique phone numbers.

or less likely to receive any particular treatment, so this restriction does not undermine the randomization. It however, selects on a set of taxpayers that on average own fewer properties: even though a unique taxpayer ID should cover every property owned by that taxpayer, in practice the TRA often issued multiple IDs for multiple properties, even when they were owned by the same person.

The actual frequencies of treatment across each group (for our set of taxpayers with unique phone numbers) are summarized in Table 2. Because we know the actual distribution of messages that were sent, we can instrument for actual message delivery with assignment to treatment.

3.4 Outcome data

Our data on outcomes was retrieved from the TRA at the beginning of August 2018. We merged the complete record of all property tax payments made for a given taxpayer ID between June and the beginning of August with our original experimental sample. For each taxpayer ID we record, for each date during this period, whether any payments associated with that ID had been made up to that date as well as the cumulative amount of payments made so far. For a subset of approximately 45,000 taxpayers we also have the final demand notice (bill) that was issued by the TRA.[7] This allows us to also investigate the impact our treatments have had on the proportion of the final tax bill each taxpayer has paid, as well as the probability that they have paid their entire bill.

4 Results

4.1 Main results

For most of the main results, we display the results from either a single treatment dummy (for any treatment) or a full set of dummies for each treatment:

$$P_i = \alpha + \beta \times treated_i + \gamma_s + \epsilon_i \tag{1}$$

and

[7] As we show in Table A2 in the Appendix, there is no imbalance within our experimental sample in the probability of being issued a bill, although for the bill subsample there are some minor imbalances between the reciprocity and social pressure treatment arms for the overall bill amount.

$$P_i = \alpha + \beta_1 T1_i + \beta_3 T2_i + \beta_4 T3_i + \gamma_s + \epsilon_i \quad (2)$$

where P_i is alternatively whether the taxpayer has paid anything to the TRA or the amount the taxpayer has paid. In equation (1), $treated_i$ is an indicator variable equal to one if the taxpayer was randomized into any of the three treatment groups. In equation (2), the indicator variables $T1_i$, $T2_i$, and $T3_i$ are dummy variables equal to one if the taxpayer was randomized into the reminder, reciprocity or social pressure treatments, respectively. Unless otherwise specified, we run both specifications (1) and (2) with strata fixed effects (indicated by γ_s) and cluster the standard errors at the taxpayer level.

Figure 1 shows the intent-to-treat (ITT) coefficient estimates from specification (1) when the outcome is whether the taxpayer has paid anything, measured at different points of time during the experiment. As can be seen, prior to the introduction of the text messages, treated taxpayers had the same propensity to have made a payment to the TRA as untreated taxpayers. Only following the introduction of the text messages do we see a difference. By the end of the period for which we have administrative data, those randomized into a message treatment were 1.8 percentage points more likely to have made a payment to the TRA, over a baseline of approximately 11 percent.

Figure 2 shows (a) the average payment rate and (b) the average amount paid for the control group and each treatment group over the course of the study relative to the timing of the treatment (indicated by the dark grey shaded bars). On average, all three treatment messages outperform the control group, with the gap opening up around the time of the first payment deadline on June 31st.

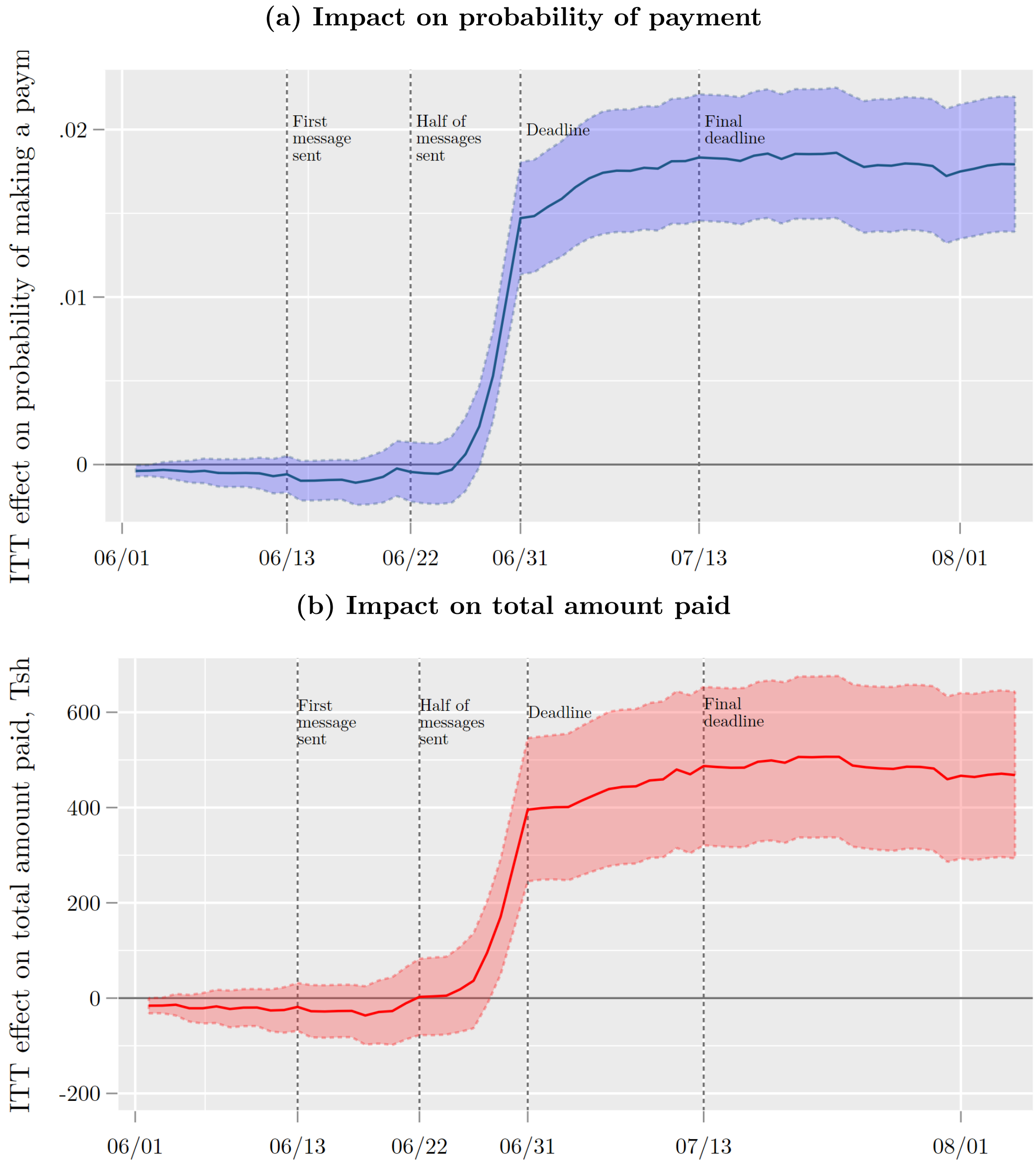
Table 3 shows the results of specification (2) when the outcome is whether the taxpayer has made any payment to the TRA measured at different points of time during the experiment. Columns (1)-(3) show the ITT estimates and (4)-(6) show the 2SLS estimates, where treatment assignment is used to instrument receipt of the correct text message. For the latter, the results indicate that one month after the deadline, those that received a reminder, reciprocity, or a social pressure message were 2, 2.3, or 1.7 percentage points more likely to make a payment, over a control mean of about 11 percentage points. At the bottom of each column we present a test of equality of these coefficients: we find that even though the reciprocity treatment has a larger point estimate than the other treatments, it only has a significantly larger impact than the social pressure treatment.

Table 4 follows the same approach, but the outcome is the amount paid in TSh. Recipients of the reminder, reciprocity, and social pressure messages paid an additional 542 TSh above a control mean of approximately 3,544 TSh, an increase in about 15 percent.[8] Those receiving the reciprocity message paid substantially more: nearly 700 TSh. This is significantly greater than reminder messages, although not significantly greater than social pressure messages at conventional levels of inference.

In Table 5 we use our subsample of taxpayers for whom we have bill data to unpack what proportion of their final tax bill has been paid, both to understand how the effects observed

[8] All TSh outcomes are winsorized at the 99th percentile unless otherwise stated. See Figure A6 in the Appendix for an analysis for how different levels of winsorizing affect the main results.

Figure 1: Timeline of ITT effect of all messages on payment rates

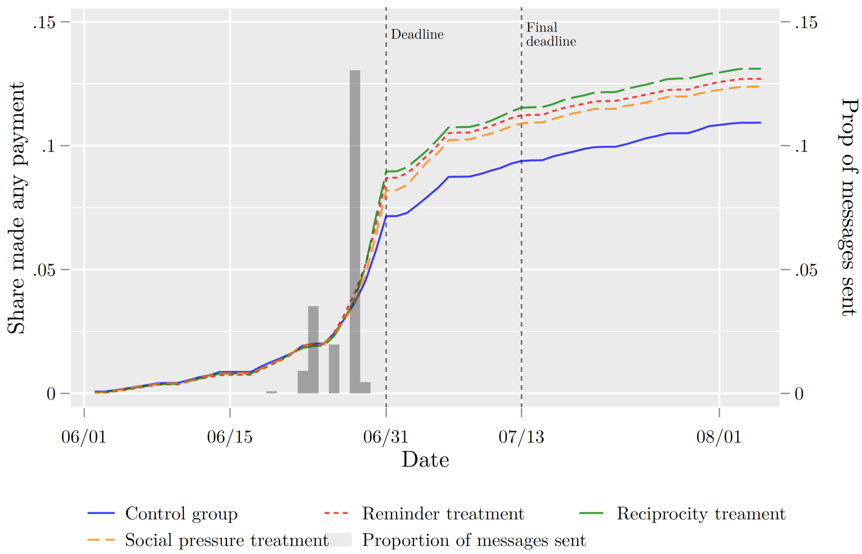


Note: Figure shows the pooled effect of being randomized into one of the treatment groups over time. Subfigure (a) shows the intent-to-treat effect on the probability that the taxpayer has made any payment to the TRA by the date shown. Subfigure (b) shows the intent-to-treat effect on the total amount the taxpayer has paid to the TRA by the date shown. All dates are for 2018. 95% confidence intervals shown.

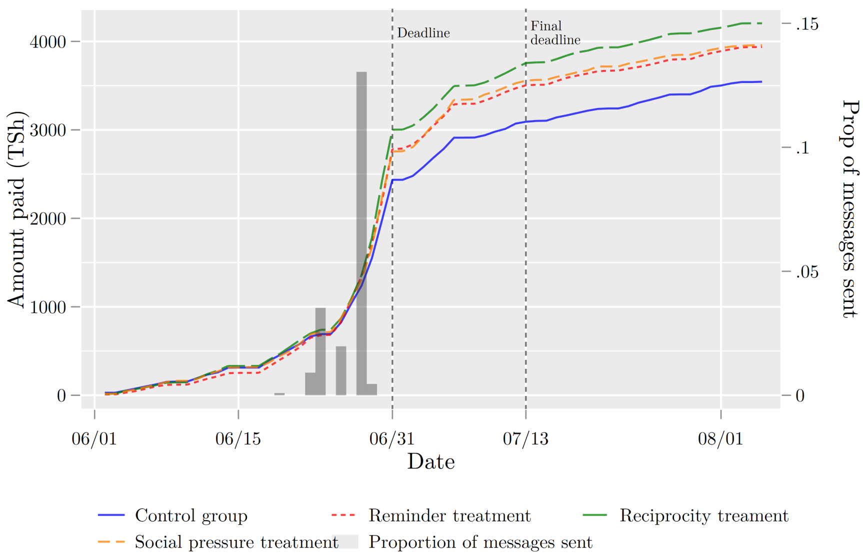
on the extensive (any payment) and intensive (how much) interact, but also to control for any differences in bill amounts between taxpayers. Column (1) replicates column (6) of Table 3, then column (2) restricts the sample to the bill subsample, showing that the effect sizes are similar across the two samples, although the difference between the reminder and other treatments shrinks. Column (3) investigates the impact of the three treatments on the proportion of the total tax bill (the 2017/18 bill and any arrears) paid by the taxpayer. During the time period we consider the control group has paid roughly 9 percent of their total bill on average, and we find that treated taxpayers pay an additional 1.6-1.9 percent of their bill on average. Only 6.9 percent of control group taxpayers paid off their entire bill. Those receiving a reminder or a

Figure 2: Timeline of payment rates by treatment group

(a) Probability of any payment to TRA



(b) Amount paid to TRA



Note: Figures show the average outcomes for each treatment group for (a) the share of taxpayers who made any payment to the TRA and (b) the average amount paid to the TRA in TSh, both over the period we have data for. The shaded bar graph indicates the timing of the treatment: each bar indicates the proportion of messages that were sent to taxpayers on a given day.

reciprocity message were 1.6-2 percentage points more likely to fully pay. Again, reciprocity performed better than the other message types, although not significantly so. This may in part be due to the narrowing of the differences between the treatments in the bill sample (moving between the first and second columns).

Table 3: Impact of message assignment on payment rates

	ITT estimates			2SLS estimates		
	Start of experiment	First tax deadline	One month after first deadline	Start of experiment	First tax deadline	One month after first deadline
Pooled treatment arms						
Treated	-0.000578	0.0147***	0.0179***	-0.000663	0.0169***	0.0205***
	(-1.04)	(8.60)	(8.72)	(-1.07)	(8.88)	(8.96)
Separate treatment arms						
T1: Reminder	-0.000738	0.0153***	0.0177***	-0.000830	0.0173***	0.0199***
	(0.000605)	(0.00190)	(0.00227)	(0.000681)	(0.00214)	(0.00256)
T2: Reciprocity	-0.000601	0.0180***	0.0220***	-0.000638	0.0191***	0.0232***
	(0.000699)	(0.00224)	(0.00267)	(0.000740)	(0.00237)	(0.00283)
T3: Social pressure	-0.000234	0.0101***	0.0144***	-0.000264	0.0117***	0.0167***
	(0.000707)	(0.00219)	(0.00263)	(0.000850)	(0.00263)	(0.00317)
Constant	0.00735***	0.0716***	0.109***			
	(0.000503)	(0.00150)	(0.00182)			
Control mean	0.007	0.072	0.109	0.007	0.072	0.109
First stage f-stat				135,432	135,432	135,432
R^2	0.013	0.041	0.048	0.000	0.001	0.001
Obs	143,425	143,425	143,425	143,425	143,425	143,425
Test: T1 = T2	0.816	0.181	0.070	0.766	0.418	0.210
Test: T1 = T3	0.401	0.008	0.156	0.428	0.016	0.254
Test: T2 = T3	0.597	0.001	0.005	0.650	0.007	0.047

Notes: Outcome is the probability a taxpayer made any payment to the TRA by the date indicated. ITT estimates indicate impact of being assigned to treatment. 2SLS estimates instrument the receipt of each message type with assignment to treatment. Robust standard-errors in parentheses. $^+p < 0.10, ^*p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

4.2 Differential impacts due to timing of treatment

As we discussed above, the three treatments will delivered to taxpayers in several sequential, overlapping batches. Many messages were received just a few days prior to the original payment deadline. This may have hampered the effectiveness of later treatments (mostly reciprocity and social pressure) if treated households had neither the time nor the resources to make a payment by the end of the month. Conversely, reminders closer to the deadline may have been *more* effective if they made the deadline more salient or reduced the chance it would be forgotten.

To check whether the timing of the messages across groups affected the impacts, we first examine whether there is any evidence of differential effects across the same treatments at different times. We do this by keeping only taxpayers in the control group and those that received a message on the five days where 99% of the messages were sent, and where there is some overlap in treatment groups (Figure A1). We then repeat our 2SLS estimation of equation (2), this time using a series of dummies capturing the interaction of the message type with the day the message was received.[9]

The results, shown in Figure A2 and Table A3, do reveal some evidence that treatment effect sizes peaked on June 25th, then declined after that (the trajectory is not fully monotonic, as messages sent before this date had a lower impact). For example, a reminder message sent on Monday, June 25th (five days before the deadline) was 32% and 120% more effective in its impact

[9]In this specification, the instrument is the type of treatment assignment interacted with the day the message was sent.

Table 4: Impact of message assignment on payment amounts

	ITT estimates			2SLS estimates		
	Start of experiment	First tax deadline	One month after first deadline	Start of experiment	First tax deadline	One month after first deadline
Pooled treatment arms						
Treated	-18.44	395.7***	468.6***	-18.77	460.1***	542.3***
	(-0.72)	(5.16)	(5.25)	(-0.66)	(5.38)	(5.45)
Separate treatment arms						
T1: Reminder	-47.06+	354.5***	401.3***	-52.97+	399.0***	451.7***
	(27.01)	(84.13)	(97.71)	(30.41)	(94.70)	(110.0)
T2: Reciprocity	12.31	559.1***	660.0***	12.96	590.6***	695.7***
	(33.94)	(100.3)	(115.8)	(35.98)	(106.4)	(122.8)
T3: Social pressure	8.130	313.8**	411.3***	9.469	362.9**	478.1***
	(33.54)	(98.08)	(114.2)	(40.32)	(117.9)	(137.2)
Constant	262.1***	2434.8***	3540.1***			
	(23.10)	(67.89)	(79.25)			
Control mean	259.641	2435.597	3544.343	259.641	2435.597	3544.343
First stage f-stat				135,432	135,432	135,432
R^2	0.019	0.047	0.055	0.000	0.000	0.000
Obs	143,425	143,425	143,425	143,425	143,425	143,425
Test: T1 = T2	0.037	0.021	0.011	0.034	0.049	0.029
Test: T1 = T3	0.050	0.638	0.920	0.064	0.725	0.824
Test: T2 = T3	0.904	0.016	0.034	0.932	0.060	0.119

Notes: Outcome is the amount in Tanzanian shillings a taxpayer paid to the TRA by the date indicated. ITT estimates indicate impact of being assigned to treatment. 2SLS estimates instrument the receipt of each message type with assignment to treatment. Payment amounts winnorsized at the 99th percentile. Robust standard-errors in parentheses. $^+p < 0.10, ^*p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

on payment rates and amounts, respectively, than one sent two days later. We see a similar effect differences for reciprocity messages as well, as well as a sharp decrease in effectiveness between social pressure messages sent on the Wednesday versus the Thursday. These results do have policy implications: messages sent at the beginning of the week before a deadline have the largest impact, or at the very least that tax authorities should not remind people only a day or two ahead of time.

To understand whether differential treatment effects across time threaten our main results, we re-run our main specifications, but this time restricting our sample to the control group and those that were sent a message on June 27th, the day when all three treatments were sent (Tables A5 and A6).[10] Our interest here is whether the relative impacts of the three treatments are similar. We find that compared to a simple reminder, the reciprocity treatment has similar or stronger relative impacts on payment rates (16% higher vs 17% originally) and amounts (90% higher vs 54% originally) as it does in the full sample, indicating that differential treatment timing may have *understated* the impact of reciprocity messages.

However, we do find that social pressure messages, which were only sent 3-4 days prior to the deadline (but only 1-2 days before the end of the working week), perform much better in relative terms when we compare them to other messages sent on the same day. In our main specification, social pressure messages performed 16% worse than reminders in payment rates and only (non-significantly) 6% better in terms of payment amounts. When comparing messages sent on the

[10]We show that our treatments are balanced within this group in Table A4.

Table 5: Payment results for subsample with bills

	Full Sample	Bill Sample		
	Pr(any payment)	Pr(any payment)	Proportion paid	Pr(fully paid)
Pooled treatment arms				
Treated	0.0205***	0.0223***	0.0183***	0.0131***
	(0.00229)	(0.00486)	(0.00422)	(0.00395)
Separate treatment arms				
T1: Reminder	0.0199***	0.0224***	0.0178***	0.0133**
	(0.00256)	(0.00545)	(0.00473)	(0.00442)
T2: Reciprocity	0.0232***	0.0241***	0.0206***	0.0155**
	(0.00283)	(0.00600)	(0.00523)	(0.00490)
T3: Social pressure	0.0167***	0.0185**	0.0146*	0.00749
	(0.00317)	(0.00676)	(0.00585)	(0.00544)
Control mean	0.109	0.109	0.089	0.069
First stage f-stat	135,432	28,065.7	28,065.7	28,065.7
R^2	0.001	0.001	0.001	0.001
Obs	143,425	31,554	31,554	31,554
Test: T1 = T2	0.210	0.752	0.564	0.633
Test: T1 = T3	0.254	0.518	0.527	0.222
Test: T2 = T3	0.047	0.418	0.313	0.152

Notes: Table shows results for payment outcomes for both entire experimental sample and the subsample for which we have bill data. The outcomes in columns (1) and(2) is an indicator equal to one if the taxpayer has made any paymen to the TRA during the study period. The outcome in column (3) is the share of the total cumulative tax bill the taxpayer made by the end of the study period. The outcome in column (4) is an indicator equal to one if the taxpayer has fully paid their tax bill by the end of the study period. All results are 2SLS results where the actual messages sent to the taxpayer are instrumented using the original treatment assignment. Robust standard-errors in parentheses. $^{+}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

same day, social pressure performed exactly as well as reminders in their impact on payment rates and 80% better in their impact on payment amounts (indistinguishable statistically from reciprocity messages). Thus, our initial estimates of the social pressure messages performing poorly is likely due to their close proximity to the deadline.

Interestingly, much of the improved performance of social pressure messages manifests *after* the initial deadline. In Figure A3, we trace out the coefficient estimates for each type of message separately, still restricting the sample to those who received a message on June 27th. As the first deadline comes to a close, social pressure messages perform about equally as well as reminder messages, but then show a much steeper improvement in payment amounts leading up to the second deadline, indicating that social pressure may have a more durable effect on compliance in the long run.

Finally, when we re-run our analysis of intensive and extensive margins from Table 5, again restricting our sample to the control group and those that were sent a message on June 27th (Table A7), we do not find any qualitative difference between our three treatments once the outcomes are transferred to be a percentage of the total bill amount. It is hard to draw firm conclusions from this, given that these results are now based on multiple subsets of the sample.

4.3 Effect sizes and cost-effectiveness

To understand the size of our effects relative to other studies, we turn to estimates published by a recent meta-analysis of tax nudges (Antinyan and Asatryan 2020), which unpack effects both on the 'extensive' margin (the propensity to pay or file) as well as the 'intensive' margin (the amount paid or reported). In our study, the pooled average effects indicate that treated taxpayers were 2 percentage points more likely to make any payment and paid roughly 15% more overall to the TRA—effects which are roughly in line with the median extensive and intensive effect sizes found for non-deterrence nudges in Antinyan and Asatryan (2020) of 1.5% and 20% respectively. When comparing our other treatments to the reminder treatment, we do not find that reciprocity and social pressure messages increase the probability of payment, which is in line with the average effects found in the meta-analysis. However, relative to reminders we do find that reciprocity-oriented messages increase the *intensive* margin of payments - the amount paid - by close to 250 TSh, or close to 7 percentage points. In Antinyan and Asatryan (2020), 'public good' oriented messages do not tend to perform better than simple reminders, indicating that on this dimension our treatment outperformed relative to what is normally observed in the literature.

Through a straightforward calculation, we estimate that the experiment's benefits are around 36 times their cost. This comes from simply dividing the average increase in the amount paid to TRA due to the intervention (around 540 TSh for those that received a message) with the text-message campaign's cost, which is estimated to be around 15 Tsh per text message. The reciprocity treatment (impact on amount paid estimated at 695 TSh) would be the most cost-effective intervention in our setting with an increase in the amount paid of more than 46 times the messages' cost.

There are, of course, several factors that this simple calculation does not consider. First, the cost of the messages could be lower if the government achieves any discount with large numbers of messages. Second, as more taxpayers comply with payment, other taxpayers could start to emulate them, and so the messages could have an additional, spillover effect that we do not consider in this simple cost-effectiveness analysis. On the other hand, we cannot rule out that the messages have only a short-term effect, although we suspect this is not the case. Previous work on text message reminders suggest that high frequency messages (e.g., daily reminders to take medicines) may lose efficacy, but not so with occasional reminders (Pop-Eleches et al. 2011).

Finally, we can also infer some insights on how program implementation affects the program's benefits. From Table 4, we observe that the ITT estimated impact is around 470 Tsh (see pooled treatment arms) while the 2SLS estimated impact is around 540 Tsh. As the ITT includes potential implementation failures (i.e., text messages not effectively received or opened), it gives an indication of the impact under imperfect implementation. On the other hand, the 2SLS gives an indication of the impact under implementation with greater fidelity. In the case of this program, implementation failures led to a roughly 13% decrease in the effectiveness of the program.

4.4 Geographic heterogeneity in treatment effects

In this subsection, we ask a simple question: had the TRA targeted a specific region of Dar es Salaam with messages, would they have had the same impact observed across the entire city? Recent meta-analyses of nudges have revealed a significant amount of heterogeneity across contexts (Antinyan and Asatryan 2020; DellaVigna and Linos 2020). This heterogeneity is important to the policymaker, who will need to know the likelihood a given treatment will be effective in a new setting and what factors might predict that effectiveness. If messages sent to taxpayers in different parts of the city have drastically different effects, then a revenue authority may want to consider targeting areas where the intervention is more cost effective, or tailor its messaging to be more effective in those areas.

As mentioned above, the randomization was conducted within 1,211 different strata, which were defined both by the location of the property (sub-ward) and whether a bill had been issued. This allows us to compare effect sizes across different locations to see if any systematic relationships appear. Many of these strata are very small in sample size, so to maximize power, we initially compare effect sizes across our three main treatments for each of the five municipalities in Dar: Kigamboni, Temeke, Ilala, Kinondoni and Ubungo. The effects, shown in Figure 3, do vary significantly across municipalities: a simple reminder message in Kigamboni is roughly 4.25 times as effective as as in Kinondoni. Despite this, the intervention is—on average—cost effective in all five municipalities. It is lowest in Kinondoni, where treated parcels remitted 166 TSh on average.

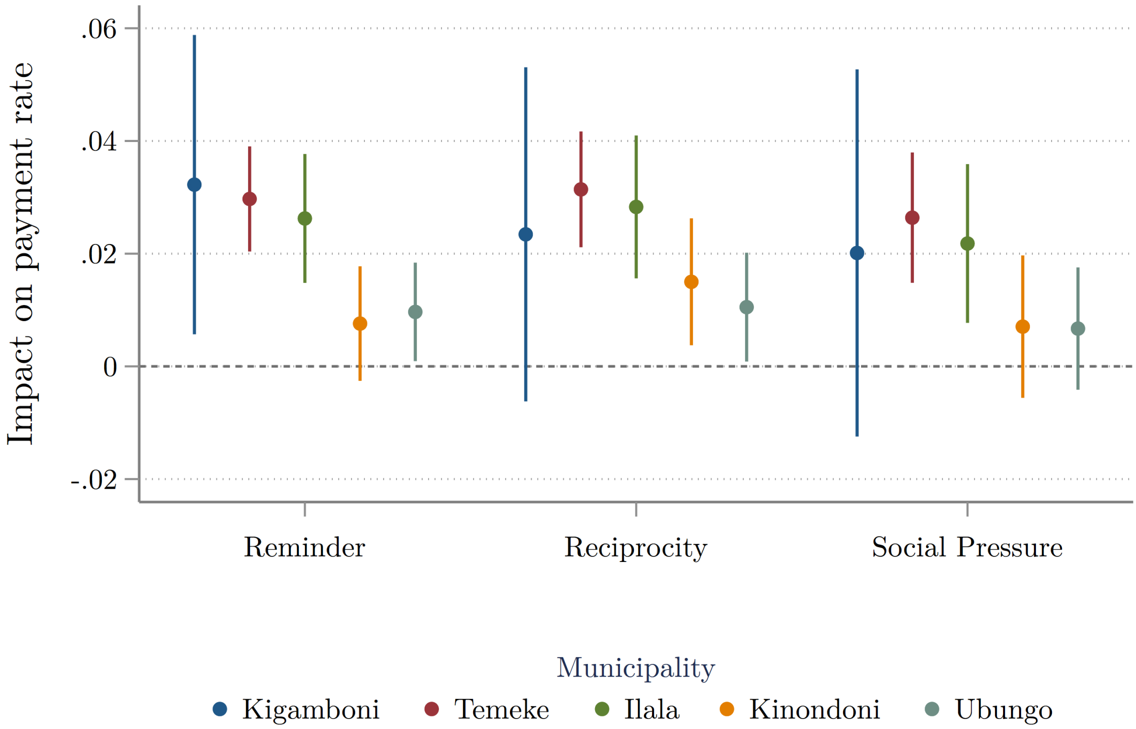
However, when we estimate the impact of receiving any message separately by ward, we find that the treatment was not cost effective in 23% of wards (Figure A4). Taking into account the cost of sending messages, had the program been extended to the entire experimental sample (i.e., including the control group), we estimate that additional revenue attributable to the message program would have been 36% higher had wards that were not cost effective been identified and screened out prior to implementation. The TRA could significantly improve the programs effectiveness if it better understood the drivers of this treatment effect heterogeneity.

To investigate whether there are any characteristics that can predict this geographic heterogeneity in the effectiveness of the text message program, we calculate average treatment on the treated (ATT) across 84 wards and also separately for 246 subwards.[11] For each geographic unit, we run our standard regressions, where the outcome is an indicator equal to one if the taxpayer made any payment to the TRA during the study period as a function of the reduced form treatment assignment. We then report the coefficient β. Because it is also of interest, we also calculate a measure of predicted control compliance, which is the presumed level of compliance that would exist in the *entire* neighborhood if the text message campaign had not taken place. We calculate this by multiplying the control group rate of compliance by the entire sample size, giving us the predicted number of taxpayers in the experimental sample that would have complied. We then add this to the number of taxpayers that had already complied before June 1st and divide by the total neighborhood population to derive the total predicted compliance rate.

To investigate neighborhood-level characteristics that might predict overall compliance and the effectiveness of the program, we implement the following process. First, we estimate the log distance of the center of the neighborhood with the nearest TRA office, under the presumption that TRA officers interact more frequently with nearby neighborhoods, and the costs of compliance are lower for taxpayers who do not have to travel far to reach a TRA office. We also

[11]For the sake of power, we drop those with fewer than 100 observations.

Figure 3: Distribution of treatment effects sizes by municipality



Note: Each coefficient is from the instrumental variable estimation of equation (2), restricting the sample to each of the five municipalities in turn. The outcome is whether the taxpayer made any payment to the TRA during the study period. When estimating equation (1) (where the treatment is receiving a message of any type), treatment effects for Temeke and Ilala are both significantly higher than Kinondoni and Ubungo at the 5% level (but not Kigamboni, due a significantly smaller sample size).

include a standardized measure of social services and infrastructure[12] and a building density measure. Finally, we include several outcomes we only observe for billed parcels. The first is the percentage of properties that are 'flat-rated', with a tax liability of 10,000 Tsh under the understanding that TRA enforcement is weaker for these areas, since ensuring compliance carries less of a return, both per parcel and overall. We also include measures of average property value multiplied by a dummy which is equal to one if there are any valued properties in the neighborhood (some subwards have no valued properties).

Table 6 displays the results of this investigation: for both our ward and sub-ward level samples, we regress both our measure of "predicted control compliance" and the estimated ATT for each neighborhood on the above set of covariates, including local authority fixed effects and (for the sub-ward samples) ward fixed effects. Predicted compliance is negatively correlated with distance to the nearest TRA office, which is consistent with closer neighborhoods facing lower costs of compliance and the TRA facing lower costs of enforcement. Building

[12]These are taken from the Resilience Academy's Climate Risk Database, which contains geocoded data specific to Dar es Salaam. For each neighborhood, we count the number of (i) hospitals, (ii) bus rapid transit stops, (iii) public water points, (iv) schools and (v) poorly managed solid waste disposal sites, each normalized by the 'population' of properties in the neighborhood. We standardize each of these outcomes (flipping the sign of outcome (v) so higher is better), then take the average of those standardized outcomes for each neighborhood.

density, correlated with both business districts and unplanned settlements in the core of the city, see higher rates of compliance. Services are either uncorrelated or negatively correlated with compliance. Finally, areas that comprise mostly flat-rated properties or do not have any valued buildings see significantly lower levels of compliance on average, consistent with our predictions that the TRA expends little effort to enforce compliance in these areas.

While we can identify several predictors of overall neighborhood compliance, very few of the neighborhood characteristics we are able to measure seem to predict the effectiveness of our treatment. There is some weak evidence that the compliance effects of the intervention were lower in subwards further away from TRA offices (column 10), but this relationship is not statistically significant in most of our results. In a separate analysis (Figure A5), we observe that the compliance rate of the control group in the experimental sample does (inversely, albeit non-linearly) predict lower levels of effectiveness, perhaps because innate drivers of compliance 'drive out' any room for impact of the treatment. We do not see a similar relationship when we use our predicted compliance measure, perhaps because this is less informative for the compliance rate of the experimental sample itself.

Although we are unable to pin down any strong geographic predictors of the effectiveness of our intervention with the data that we have, governments and researchers with geographically-explicit data may be better positioned to understand such heterogeneity in future work.

4.5 Distributional effects and individual-level heterogeneity

Another dimension we explore in the data is the distribution of effects across types of properties. For the sample of properties for which a bill was issued, we order the bills by amount owed percentiles (for the 17/18 tax year, not including back taxes) and plot the percentage of the total potential bill revenue for each percentile in Figure 4. The graph indicates that in terms of potential revenues, most of the opportunity for increased revenue would come from the top 10 percent of billed properties, which account for more than 80 percent of the total potential bill revenue. We then look at treatment heterogeneity across a property's bill status (was a bill issued?) and then, for properties that have received a bill, whether treatment effects are different for (i) properties that were flat-rated (only receiving a TSh 10,000 bill with no attempt to value the property) and (ii) for different bill amounts, dividing the sample into four subsets roughly equivalent to quintiles.[13] Table A1 displays the results. There is little evidence of substantial heterogeneity across bill amounts, only some evidence that reminder messages had a weakly-stronger effect for those who received bill amounts between TSh 11-25,000. We thus cannot draw any strong conclusions about the progressivity of the treatment, although it does appear that in aggregate, where a taxpayer resides in the city appears to have stronger explanatory power for the success of the treatment than that taxpayer's bill size, holding location constant.

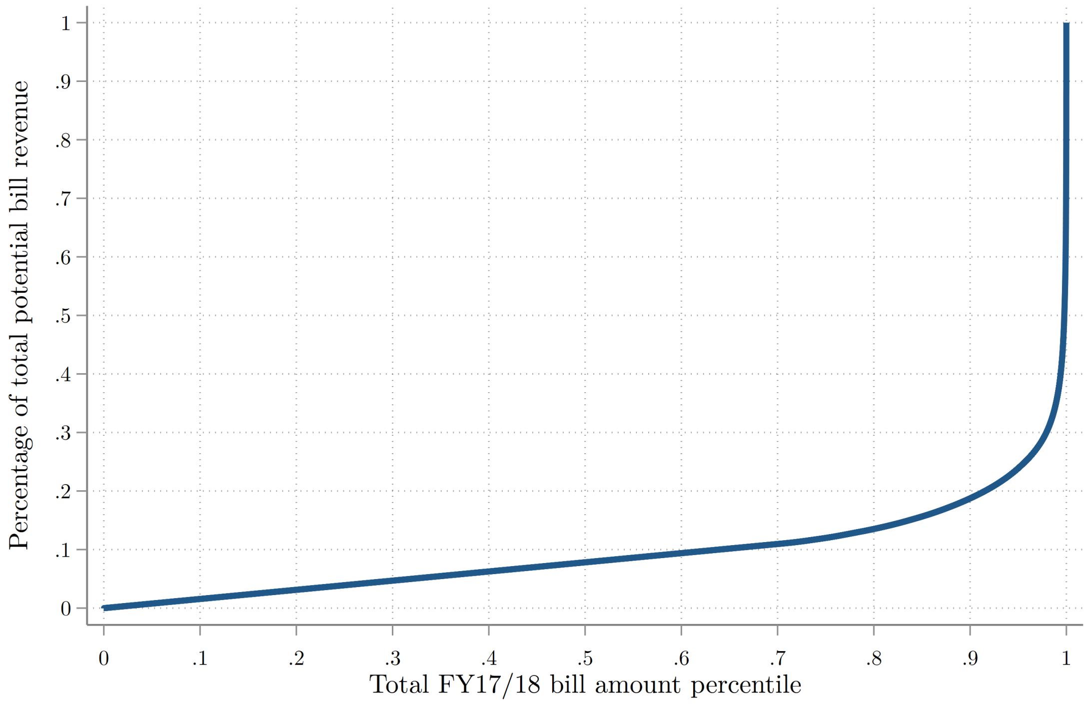
[13]Because many taxpayers received the same bill amount (e.g. TSh 10,000 or 25,000, it is not possible to perfectly divide the sample into quintiles or deciles)

Table 6: Predictors of geographic heterogeneity in overall compliance and intervention treatment effects

	Outcome: Predicted neighborhood compliance without intervention					Outcome: Average treatment effect in experiment				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Log(Distance to nearest TRA office)	-0.0815***	-0.0648***	-0.0515***	-0.0579***	0.00656	-0.0108	-0.00577	-0.00114	0.00165	-0.0272*
	(0.0120)	(0.0119)	(0.0123)	(0.0128)	(0.0282)	(0.00911)	(0.00902)	(0.00676)	(0.00672)	(0.0132)
Number of buildings per sq km	17.91^{+}	30.32***	27.03***	15.68*	13.20	-4.637	-6.206	-0.293	-1.670	-7.516^{+}
	(9.516)	(7.864)	(7.282)	(6.472)	(9.147)	(5.186)	(5.202)	(1.669)	(1.823)	(4.162)
Standardized index of services	-0.00535	-0.0207	-1.536	-1.927*	-1.923	0.00653	0.00269	0.127	0.295	0.0325
	(0.0191)	(0.0169)	(0.948)	(0.969)	(1.173)	(0.00975)	(0.0104)	(0.274)	(0.283)	(0.433)
% of billed parcels that are flatrated	-0.0685*	-0.204***	-0.153**	-0.306***	-0.336***	0.00699	-0.0325	0.0263	-0.0792^{+}	-0.0230
	(0.0315)	(0.0570)	(0.0497)	(0.0575)	(0.0826)	(0.0202)	(0.0258)	(0.0203)	(0.0412)	(0.0296)
Any valued buildings?			0.476**	0.413*	0.0680			-0.0653	-0.110*	-0.0295
			(0.172)	(0.160)	(0.164)			(0.0507)	(0.0558)	(0.0616)
Log(mean building value)	-0.0339**	-0.0269**	-0.0176^{+}	-0.0120	0.00712	0.00804	0.00572	0.00335	-0.00171	0.00194
	(0.0107)	(0.00906)	(0.0105)	(0.00956)	(0.0102)	(0.00635)	(0.00623)	(0.00321)	(0.00390)	(0.00348)
Constant	2.100***	1.937***	0.941***	1.147***	0.688*	-0.0210	-0.00450	0.0283	0.112^{+}	0.275*
	(0.237)	(0.176)	(0.123)	(0.127)	(0.272)	(0.147)	(0.138)	(0.0641)	(0.0655)	(0.124)
Sample	ward	ward	subward	subward	subward	ward	ward	subward	subward	subward
R^2	0.552	0.737	0.288	0.452	0.669	0.103	0.211	0.011	0.078	0.466
Obs	84	84	246	246	246	84	84	246	246	246
Ward f.e.					X					X
Local authority f.e.		X		X			X		X	

Notes: Outcome for columns (1-5) is the predicted compliance of the neighborhood (subward or ward) absent the intervention. This is calculated as the control group compliance rate multiplied by the number of properties in the experimental sample (those that had not paid before June 1st), added to the total number of properties that had paid before June 1st (not part of the experimental sample), divided by the total population of properties in the neighbordhood. The outcome for columns (6-10) is the coefficient estimate of the neighbordhood-specific treatment effect for receiving any message. Covariates include: the log of the distance (in kilometers) to the nearest TRA office, building density, a standardized index of the presence of social services (hospitals, schools, bus rapid transit stops, public water points, and a lack of poorly-managed solid waste sites), the % of billed parcels that are at 10,000 TSh, and a Gini Index measure of inequality of property valuations. Robust standard-errors in parentheses. $^{+}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Figure 4: Potential bill revenue and bill amount percentiles



Note: Figure shows (on the y-axis) the cumulative proportion of the total the total 2017/18 fiscal year bill amount charged by the TRA against (on the x-axis) the percentile of bill amounts the taxpayer is in. For example, the first 50% of taxpayers, ranked by their bill amount, account for last than 5% being charged by the TRA.

5 Conclusion

In the face of limited tax compliance and limited capacity for enforcement of tax compliance, this study reports the impact of a randomized controlled trial to test different ways of leveraging tax morale in Dar es Salaam, Tanzania. The study tests salience (via a simple reminder), reciprocity (via a message highlighting the link between tax payment and publicly provided services), and social pressure (via a message emphasizing that those who do not pay are not contributing to national or local development).

We find positive impacts of all the messages, suggesting an impact of simple salience (since any message boosts salience). But we find evidence that reciprocity treatments led to higher overall tax payments, suggesting that it is possible for government to leverage these aspects of tax morale. While the absolute gains are not enormous, the interventions are very cheap, such that the benefit to cost ratio is estimated to be over 35:1 on average across all treatments and over 45:1 for the most effective treatment. While it is not clear whether these nudges would have long term impacts on tax compliance (something we were unable to test in this setting), they demonstrate that working with tax authorities to design experiments can relatively quickly provide evidence that can be used to adjust tax administration practices in the short term.

We conclude that text message reminders are one useful tool that governments can draw on to mobilize domestic resources for public services. To substantively raise domestic revenues in absolute terms, tax systems will require a broad range of improvements, including more expansive registration and valuation and more effective enforcement. Future work in this area can be designed and statistically powered to test the progressivity of tax morale interventions and a wider range of potential messages to ascertain the sensitivity of tax morale interventions to implementation details around wording, length, and frequency.

Data availability statement

The tax data used in this paper were received from the TRA for the purposes of this research, and because this analysis uses the trajectory of payments, anonymization of the data is not straightforward. Thus, the data are not publicly available. We would be happy to run any requested sensitivity analysis, and we welcome other researchers to approach the TRA for access to revenue data.

References

Ali, M., O.-H. Fjeldstad, and L. Katera (2017). Property taxation in developing countries. *CMI Brief*.

Allingham, M. G. and A. Sandmo (1972). Income tax evasion: a theoretical analysis. *Journal of Public Economics 1*(3-4), 323–338.

Antinyan, A. and Z. Asatryan (2020). Nudging for tax compliance: A meta-analysis. *cesifo Working Papers 8500.*

Ariel, B. (2012). Deterrence and moral persuasion effects on corporate tax compliance: findings from a randomized controlled trial. *Criminology 50*(1), 27–69.

Bergeron, A., P. Bessone, J. K. Kabeya, G. Z. Tourek, and J. L. Weigel (2022). Optimal assignment of bureaucrats: Evidence from randomly assigned tax collectors in the drc. Technical report, National Bureau of Economic Research.

Besley, T. and T. Persson (2013). Taxation and development. In *Handbook of public economics*, Volume 5, pp. 51–110. Elsevier.

Blumenthal, M., C. Christian, J. Slemrod, and M. G. Smith (2001). Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota. *National Tax Journal*, 125–138.

Brockmeyer, A., A. M. Estefan, K. R. Arras, and J. C. S. Serrato (2020). Taxing property in developing countries: Theory and evidence from mexico. *Working Paper*.

Brockmeyer, A., S. Smith, M. Hernandez, and S. Kettle (2019). Casting a wider tax net: Experimental evidence from Costa Rica. *American Economic Journal: Economic Policy 11*(3), 55–87.

Carril, A. et al. (2017). Dealing with misfits in random treatment assignment. *Stata Journal 17*(3), 652–667.

Castro, L. and C. Scartascini (2015). Tax compliance and enforcement in the Pampas evidence from a field experiment. *Journal of Economic Behavior & Organization 116*, 65–82.

Cohen, I. (2020). Low-cost tax capacity: A randomized evaluation with the Uganda Revenue Authority.

Coleman, S. (1996). The Minnesota income tax compliance experiment.

Del Carmen, G., E. E. Espinal Hernandez, and T. Scot (2022). Targeting in tax compliance interventions: Experimental evidence from honduras. *Working Paper*.

Del Carpio, L. (2022). Are the neighbors cheating? evidence from a social norm experiment on property taxes in peru. Technical report, INSEAD.

DellaVigna, S. and E. Linos (2020). Rcts to scale: Comprehensive evidence from two nudge units. Technical report, National Bureau of Economic Research.

Dincecco, M. and G. Katz (2014). State capacity and long-run economic performance. *The Economic Journal 126*(590), 189–218.

Fellner, G., R. Sausgruber, and C. Traxler (2013). Testing enforcement strategies in the field: Threat, moral appeal and social information. *Journal of the European Economic Association 11*(3), 634–660.

Fjeldstad, O.-H., M. Ali, and L. Katera (2017). Taxing the urban boom in Tanzania: Central versus local government property tax collection. *CMI Insight*.

Gadenne, L. (2017). Tax me, but spend wisely? sources of public finance and government accountability. *American Economic Journal: Applied Economics*.

Gaspar, V., L. Jaramillo, and M. P. Wingender (2016). *Tax Capacity and Growth: Is there a Tipping Point?* International Monetary Fund.

Hallsworth, M., J. A. List, R. D. Metcalfe, and I. Vlaev (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics 148*, 14–31.

Hasseldine, J., P. Hite, S. James, and M. Toumi (2007). Persuasive communications: Tax compliance enforcement strategies for sole proprietors. *Contemporary Accounting Research 24*(1), 171–194.

Hernandez, M., J. Jamison, E. Korczyc, N. Mazar, and R. Sormani (2017). Applying behavioral insights to improve tax collection: experimental evidence from Poland.

Herrmann, B., C. Toni, and S. Gachter (2008). Antisocial punishment across societies. *Science 319*, 1362–167.

Holz, J. E., J. A. List, A. Zentner, M. Cardoza, and J. E. Zentner (2023). The $100 million nudge: Increasing tax compliance of firms using a natural field experiment. *Journal of Public Economics 218*, 104779.

Hoy, C., M. Sinning, and L. McKenzie (2022). Improving tax compliance without increasing revenue: Evidence from population-wide randomized controlled trials in papua new guinea. *Economic Development and Cultural Change*.

Kettle, S., M. Hernandez, S. Ruda, and M. Sanders (2016). *Behavioral interventions in tax compliance: evidence from Guatemala*. The World Bank.

Khwaja, A. I., O. Haq, A. Q. Khan, B. Olken, and M. Shaukat (2020). *Rebuilding the social compact: Urban service delivery and property taxes in Pakistan*. 3ie.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in Denmark. *Econometrica 79*(3), 651–692.

Krause, B. (2020). Balancing purse and peace: Tax collection, public goods and protests.

Luttmer, E. F. and M. Singhal (2014). Tax morale. *Journal of Economic Perspectives 28*, 149–168.

Mascagni, G. and C. Nell (2022). Tax compliance in Rwanda: Evidence from a message field experiment. *Economic Development and Cultural Change 70*(2), 587–623.

Meiselman, B. S. (2018). Ghostbusting in Detroit: Evidence on nonfilers from a controlled field experiment. *Journal of Public Economics 158*, 180–193.

Mwaijande, F., M. Kachwamba, J. Mwakalikamo, D. Shirima, and G. Cruces (2021). Local authorities and tax collection: Experimental evidence from Tanzania.

Okunogbe, O. and F. Santoro (2023). Increasing tax collection in african countries: The role of information technology. *Journal of African Economies 32*(Supplement_1), i57–i83.

Ortega, D. and C. Scartascini (2020). Don't blame the messenger. the delivery method of a message matters. *Journal of Economic Behavior & Organization 170*, 286–300.

PO-RALG (2013). A study on LGAs own source revenue collection. *PO-RALG - Internal Report*.

Pomeranz, D. (2015). No taxation without information: Deterrence and self-enforcement in the value added tax. *American Economic Review 105*(8), 2539–69.

Pop-Eleches, C., H. Thirumurthy, J. P. Habyarimana, J. G. Zivin, M. P. Goldestein, D. de Walque, L. MacKeen, J. Haberer, S. Kimaiyo, J. Sidle, D. Ngare, and D. R. Bangsberg (2011). Mobile phone technologies improve adherence to antiretroviral treatment in a resource-limited setting: a randomized controlled trial of text message reminders. *AIDS 25*(6), 825–834.

Santoro, F. and R. Waiswa (2023). How to improve tax compliance by wealthy persons? evidence from uganda. *Development Policy Review*, e12754.

Shimeles, A., D. Z. Gurara, and F. Woldeyes (2017). Taxman's dilemma: coercion or persuasion? evidence from a randomized field experiment in ethiopia. *American Economic Review 107*(5), 420–424.

Torgler, B. (2004). Moral suasion: An alternative tax policy strategy? Evidence from a controlled field experiment in Switzerland. *Economics of Governance 5*(3), 235–253.

A1 Appendix

A1.1 Additional Graphs and Tables

Table A1: Heterogeneity in message effectiveness across bill amounts

	Full Sample	Bill Sample			
	Pr(any payment)	Pr(any payment)	Pr(any payment)	Proportion paid	Proportion paid
Pooled treatment arms					
Treated	0.0195***	0.0203*	0.0185*	0.0142	0.0167
	(0.00261)	(0.00921)	(0.00898)	(0.00938)	(0.00952)
Treated × billed	0.00243				
	(0.00551)				
Treated × flat-rated		0.00434		0.00828	
		(0.0114)		(0.0116)	
Treated × bill(11-25k)			0.0131		0.00574
			(0.0118)		(0.0123)
Treated × bill(25-70k)			0.00135		-0.00846
			(0.0153)		(0.0160)
Treated × bill(70k and above)			-0.00719		-0.000284
			(0.0148)		(0.0155)
Separate treatment arms					
Reminder	0.0192***	0.0208*	0.0159	0.0189*	0.0137
	(0.00290)	(0.00914)	(0.00991)	(0.00934)	(0.0105)
Reciprocity	0.0229***	0.0289**	0.0218*	0.0216*	0.0228+
	(0.00321)	(0.0101)	(0.0111)	(0.0102)	(0.0116)
Social Pressure	0.0162***	0.0160	0.0203	0.00421	0.0166
	(0.00359)	(0.0113)	(0.0127)	(0.0114)	(0.0133)
Reminder × billed	0.00317				
	(0.00614)				
Reciprocity × billed	0.00106				
	(0.00678)				
Social Pressure × billed	0.00240				
	(0.00762)				
Reminder × flat-rated		0.00395		0.00298	
		(0.0112)		(0.0115)	
Reciprocity × flat-rated		-0.00772		-0.00140	
		(0.0124)		(0.0126)	
Social Pressure × flat-rated		0.00510		0.0147	
		(0.0139)		(0.0141)	
Reminder × bill(11-25k)			0.0247+		0.0198
			(0.0132)		(0.0137)
Reminder × bill(25-70k)			0.000228		-0.00768
			(0.0170)		(0.0178)
Reminder × bill(70k and above)			-0.0133		-0.00512
			(0.0164)		(0.0172)
Reciprocity × bill(11-25k)			0.00198		-0.00735
			(0.0146)		(0.0151)
Reciprocity × bill(25-70k)			0.0114		-0.00613
			(0.0188)		(0.0196)
Reciprocity × bill(70k and above)			-0.00389		0.00362
			(0.0180)		(0.0188)
Social Pressure × bill(11-25k)			0.000488		-0.00996
			(0.0166)		(0.0170)
Social Pressure × bill(25-70k)			-0.00880		-0.0135
			(0.0211)		(0.0221)
Social Pressure × bill(70k and above)			0.000318		0.00409
			(0.0204)		(0.0212)
Control mean					
First stage f-stat					
R^2	0.001	0.004	0.005	0.032	0.014
Obs	143,425	31,554	31,554	31,554	31,554

Notes: Table shows results for payment outcomes for both entire experimental sample and the subsample for which we have bill data. The outcomes in columns (1) and(2) is an indicator equal to one if the taxpayer has made any paymen to the TRA during the study period. The outcome in columns (3-5) is the share of the total cumulative tax bill the taxpayer made by the end of the study period. Billed = taxpayers that received a bill. Flat-rated are billed taxpayers that were assigned a flat rate of 10,000 Tsh. bill(11-25k) indicates the billed taxpayer received a bill between those amongs, and so on (the leave-out category are those that had a cumulative bill of 10,000 or below). All results are 2SLS results where the actual messages sent to the taxpayer and their interactions are instrumented using the original treatment assignment and its interaction. Robust standard-errors in parentheses. $^{+}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Table A2: Balance tests for bill subsample

	Full Sample	Bill Sample			
	(1) Probability(bill issued)	(2) Log(total bill)	(3) Log(17/18 bill)	(4) IHS(property value)	(5) Flat-rated
T1: Reminder	-0.00164	-0.0107	-0.00848	-0.128	0.00760
	(0.00300)	(0.0132)	(0.0116)	(0.0907)	(0.00510)
T2: Reciprocity	-0.00112	-0.00722	0.00317	-0.0763	0.00466
	(0.00347)	(0.0152)	(0.0135)	(0.105)	(0.00588)
T3: Social pressure	-0.00316	0.00816	0.00631	0.00126	0.000476
	(0.00347)	(0.0154)	(0.0136)	(0.106)	(0.00594)
Control mean (tsh)	.2222144628	268,066.239	120,048.821	74,139,772	.6381206788
R^2	0.000	0.465	0.549	0.547	0.535
Obs	143,520	31,550	27,090	31,554	31,554
Test: T1 = T2	0.861	0.789	0.307	0.570	0.562
Test: T1 = T3	0.612	0.152	0.202	0.159	0.167
Test: T2 = T3	0.554	0.312	0.815	0.461	0.480

Notes: Table shows balance tests for four bill related outcomes. Outcome (1) is a binary outcome = 1 if the taxpayer was issued a bill by the TRA. Outcome (2) is the log of the cumulative bill amount (current and back taxes) across all the taxpayer's properties. Outcome (3) is the log of the 2017/2018 tax owed only. Outcome (4) is the inverse hyperbolic sign transformation of the assessed property value by the TRA. The treatment measures are the original, intent-to-treat indicators. Robust standard-errors in parentheses. $^{+}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Table A3: Impacts of receiving the message on different dates: regression results

	Outcome = payment rate	Outcome = payment amount
Reminder (June 23rd)	0.0233***	670.4***
	(8.34)	(5.59)
Reminder (June 25th)	0.0260***	801.9**
	(4.32)	(3.10)
Reminder (June 27th)	0.0196***	360.9**
	(7.52)	(3.29)
Reminder (June 22nd)	0.0197***	249.0
	(3.56)	(1.12)
Reciprocity (June 27th)	0.0222***	613.5***
	(7.45)	(4.80)
Reciprocity (June 25th)	0.0298***	977.9***
	(7.81)	(5.82)
Social Pressure (June 27th)	0.0196***	538.0***
	(7.00)	(4.48)
Social Pressure (June 28th)	0.0135*	306.4
	(2.25)	(1.17)
First stage f-stat	36,555.8	36,555.8
R^2	0.001	0.000
Obs	143,025	143,025
Test: Reminder(23) = Reminder(27)	0.231	0.020
Test: Reciprocity(25) = Reprocity(27)	0.082	0.055
Test: Pressure(27) = Pressure(28)	0.327	0.394

Notes: Outcome is the probability a taxpayer made any payment to the TRA or the amount (in TSh) made. 2SLS estimates instrument the receipt of each message type × the receipt date with assignment to treatment × receipt date. Taxpayers who received any messages before these dates are dropped from the specification. Robust standard-errors in parentheses. $^{+}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Table A4: Balance tests for bill subsample (June 27th sample)

	Full Sample	Bill Sample			
	(1) Probability(bill issued)	(2) Log(total bill)	(3) Log(17/18 bill)	(4) IHS(property value)	(5) Flat-rated
T1: Reminder	-0.00407	-0.0210	-0.00943	-0.173$^+$	0.0104$^+$
	(0.00341)	(0.0150)	(0.0132)	(0.103)	(0.00581)
T2: Reciprocity	-0.00126	-0.0193	-0.000368	-0.0182	0.000596
	(0.00386)	(0.0170)	(0.0151)	(0.117)	(0.00657)
T3: Social pressure	-0.00465	0.00347	0.00589	-0.0418	0.00296
	(0.00357)	(0.0160)	(0.0142)	(0.110)	(0.00617)
Control mean (tsh)	.2217941084	268,436.619	119,837.641	74,092,386	.6383449147
R^2	0.000	0.465	0.544	0.552	0.540
Obs	103,718	22,618	19,435	22,621	22,621
Test: T1 = T2	0.461	0.918	0.534	0.176	0.129
Test: T1 = T3	0.869	0.113	0.262	0.221	0.218
Test: T2 = T3	0.391	0.191	0.685	0.845	0.728

Notes: Table shows balance tests for four bill related outcomes. Outcome (1) is a binary outcome = 1 if the taxpayer was issued a bill by the TRA. Outcome (2) is the log of the cumulative bill amount (current and back taxes) across all the taxpayer's properties. Outcome (3) is the log of the 2017/2018 tax owed only. Outcome (4) is the inverse hyperbolic sign transformation of the assessed property value by the TRA. The treatment measures are the original, intent-to-treat indicators. Sample is restricted to control group or taxpayers who were sent a message on June 27th. Robust standard-errors in parentheses. $^+p < 0.10, ^* p < 0.05, ^{**} p < 0.01, ^{***} p < 0.001$

Table A5: Impact of message assignment on payment rates for taxpayers that only received a message on June 27th

	ITT estimates			2SLS estimates		
	Start of experiment	First tax deadline	One month after first deadline	Start of experiment	First tax deadline	One month after first deadline
Pooled treatment arms						
Treated	-0.000528	0.0115***	0.0155***	-0.000555	0.0140***	0.0187***
	(-0.90)	(6.40)	(7.13)	(-0.80)	(6.57)	(7.30)
Separate treatment arms						
T1: Reminder	-0.000827	0.0108***	0.0139***	-0.00105	0.0137***	0.0176***
	(0.000683)	(0.00217)	(0.00260)	(0.000863)	(0.00274)	(0.00328)
T2: Reciprocity	-0.000252	0.0149***	0.0190***	-0.000260	0.0160***	0.0204***
	(0.000789)	(0.00249)	(0.00298)	(0.000864)	(0.00274)	(0.00327)
T3: Social pressure	-0.000382	0.00988***	0.0147***	-0.000468	0.0118***	0.0176***
	(0.000724)	(0.00226)	(0.00272)	(0.000893)	(0.00279)	(0.00336)
Constant	0.00733***	0.0715***	0.109***			
	(0.000504)	(0.00151)	(0.00182)			
Control mean	0.007	0.071	0.109	0.007	0.071	0.109
First stage F-stat				74,189.2	74,189.2	74,189.2
R^2	0.015	0.044	0.050	0.000	0.001	0.001
Obs	103,616	103,616	103,616	103,616	103,616	103,616
Test: T1 = T2	0.449	0.112	0.088	0.382	0.434	0.423
Test: T1 = T3	0.522	0.675	0.774	0.509	0.498	0.994
Test: T2 = T3	0.871	0.056	0.164	0.832	0.179	0.460

Notes: Outcome is the probability a taxpayer made any payment to the TRA by the date indicated. ITT estimates indicate impact of being assigned to treatment. 2SLS estimates instrument the receipt of each message type with assignment to treatment. Robust standard-errors in parentheses. Sample restricted to taxpayers that received no message or only received a message on June 27th. $^+p < 0.10, ^* p < 0.05, ^{**} p < 0.01, ^{***} p < 0.001$

Table A6: Impact of message assignment on payment amounts for taxpayers that only received a message on June 27th

	ITT estimates			2SLS estimates		
	Start of experiment	First tax deadline	One month after first deadline	Start of experiment	First tax deadline	One month after first deadline
Pooled treatment arms						
Treated	-10.34	289.2***	378.8***	-4.865	367.9***	474.6***
	(-0.38)	(3.60)	(4.05)	(-0.15)	(3.87)	(4.29)
Separate treatment arms						
T1: Reminder	-47.98	177.6+	233.5*	-60.71	224.5+	295.3*
	(30.32)	(94.83)	(110.2)	(38.35)	(120.0)	(139.5)
T2: Reciprocity	20.16	442.2***	527.7***	21.88	477.0***	565.1***
	(38.45)	(111.2)	(128.1)	(42.20)	(122.0)	(140.4)
T3: Social pressure	11.39	305.8**	438.8***	13.41	364.6**	527.6***
	(34.85)	(100.8)	(117.6)	(43.05)	(124.4)	(145.2)
Constant	261.7***	2428.3***	3525.1***			
	(23.20)	(67.77)	(79.10)			
Control mean	259.058	2422.307	3522.376	259.058	2422.307	3522.376
First stage F-stat				74,189.2	74,189.2	74,189.2
R^2	0.022	0.049	0.057	0.000	0.000	0.000
Obs	103,616	103,616	103,616	103,616	103,616	103,616
Test: T1 = T2	0.061	0.016	0.020	0.052	0.052	0.071
Test: T1 = T3	0.070	0.199	0.077	0.072	0.265	0.111
Test: T2 = T3	0.827	0.237	0.503	0.864	0.427	0.818

Notes: Outcome is the amount in Tanzanian shillings a taxpayer paid to the TRA by the date indicated. ITT estimates indicate impact of being assigned to treatment. 2SLS estimates instrument the receipt of each message type with assignment to treatment. Sample restricted to taxpayers that received no message or only received a message on June 27th. Payment amounts winnorsized at the 99th percentile. Robust standard-errors in parentheses. $^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table A7: Payment results for subsample with bills and those that only received a message on June 27th

	Full Sample	Bill Sample		
	Pr(any payment)	Pr(any payment)	Proportion paid	Pr(fully paid)
Pooled treatment arms				
Treated	0.0187***	0.0222***	0.0197***	0.0171***
	(0.00256)	(0.00547)	(0.00495)	(0.00470)
Separate treatment arms				
T1: Reminder	0.0176***	0.0235***	0.0207**	0.0192**
	(0.00328)	(0.00709)	(0.00641)	(0.00610)
T2: Reciprocity	0.0204***	0.0226**	0.0199**	0.0188**
	(0.00327)	(0.00698)	(0.00632)	(0.00603)
T3: Social pressure	0.0176***	0.0204**	0.0186**	0.0132*
	(0.00336)	(0.00721)	(0.00652)	(0.00617)
Control mean	0.109	0.108	0.094	0.078
First stage f-stat	74,189.2	15,655.3	15,655.3	15,655.3
R^2	0.001	0.001	0.001	0.001
Obs	103,616	22,621	22,621	22,621
Test: T1 = T2	0.423	0.909	0.910	0.954
Test: T1 = T3	0.994	0.680	0.758	0.352
Test: T2 = T3	0.460	0.787	0.860	0.428

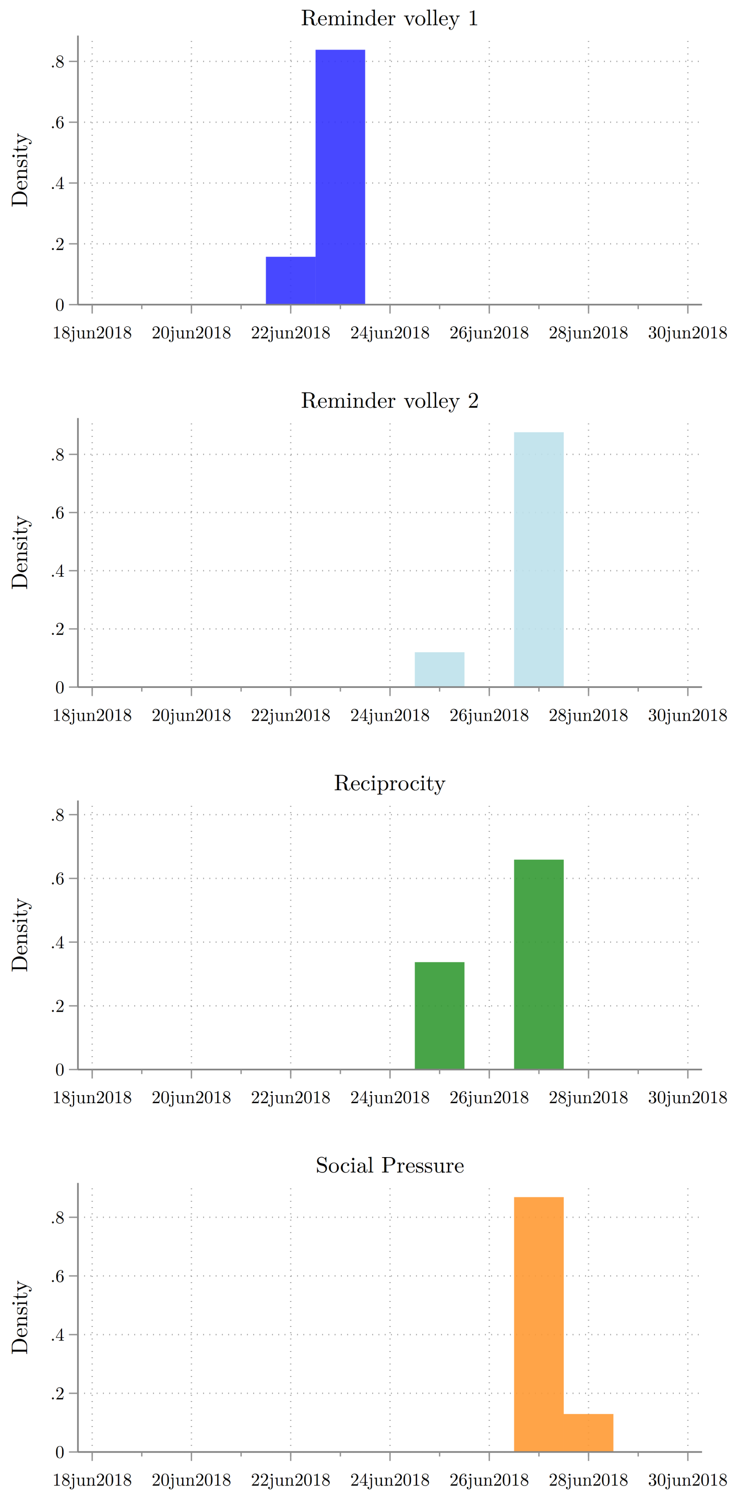
Notes: Table shows results for payment outcomes for both entire experimental sample and the subsample for which we have bill data. The outcomes in columns (1) and(2) is an indicator equal to one if the taxpayer has made any paymen to the TRA during the study period. The outcome in column (3) is the share of the total cumulative tax bill the taxpayer made by the end of the study period. The outcome in column (4) is an indicator equal to one if the taxpayer has fully paid their tax bill by the end of the study period. All results are 2SLS results where the actual messages sent to the taxpayer are instrumented using the original treatment assignment. Robust standard-errors in parentheses. $^{+}p < 0.10, ^{*} p < 0.05, ^{**} p < 0.01, ^{***} p < 0.001$

Table A8: Test of difference of pre-existing covariates for taxpayers who received messages on different dates

	Full Sample	Bill Sample			
	(1)	(2)	(3)	(4)	(5)
	Probability(bill issued)	Log(total bill)	Log(17/18 bill)	IHS(property value)	Flat-rated
Message sent on:					
June 22nd	0.000902	-0.0590+	-0.0492	-0.240	0.0113
	(0.000867)	(0.0347)	(0.0301)	(0.241)	(0.0136)
June 23rd	0.000592	-0.0298	-0.0233	-0.0573	0.00147
	(0.000529)	(0.0205)	(0.0179)	(0.142)	(0.00802)
June 27th	0.000425	-0.0457**	-0.0197	-0.0722	0.00236
	(0.000454)	(0.0176)	(0.0154)	(0.123)	(0.00690)
June 28th	0.000121	0.0391	0.0212	0.446+	-0.0261+
	(0.000921)	(0.0351)	(0.0305)	(0.244)	(0.0138)
Constant	0.220***	10.36***	9.799***	6.510***	0.638***
	(0.000414)	(0.0161)	(0.0140)	(0.112)	(0.00629)
R^2					
Obs	.9876418	.4692811	.5591222	.5525018	.5402233
N	102,545	22,489	19,278	22,491	22,491

Notes: Table shows results from a linear regression of different balance outcomes on dummies indicating which date the taxpayer received a message. The omitted date is June 25th, which, as indicated in in Figure A2, was the date where messages were shown to have the largest impact. The sample is restricted to only treated taxpayers. Table shows balance tests for four bill related outcomes. Outcome (1) is a binary outcome = 1 if the taxpayer was issued a bill by the TRA. Outcome (2) is the log of the cumulative bill amount (current and back taxes) across all the taxpayer's properties. Outcome (3) is the log of the 2017/2018 tax owed only. Outcome (4) is the inverse hyperbolic sign transformation of the assessed property value by the TRA. (5) Is an indicator if the taxpayer was billed and flatrated (issued a bill for a flat 10,000 TSh) Robust standard-errors in parentheses. $^{+}p < 0.10, ^{*} p < 0.05, ^{**} p < 0.01, ^{***} p < 0.001$

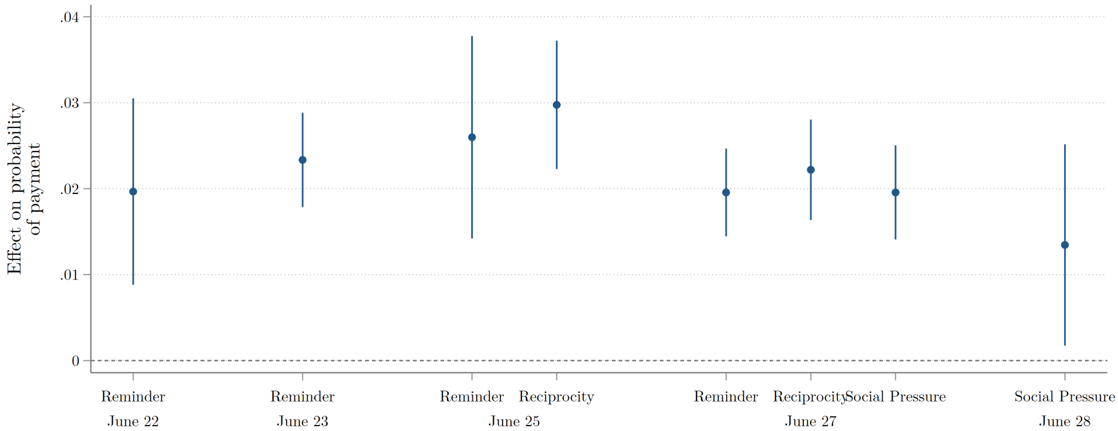
Figure A1: Timeline of message delivery



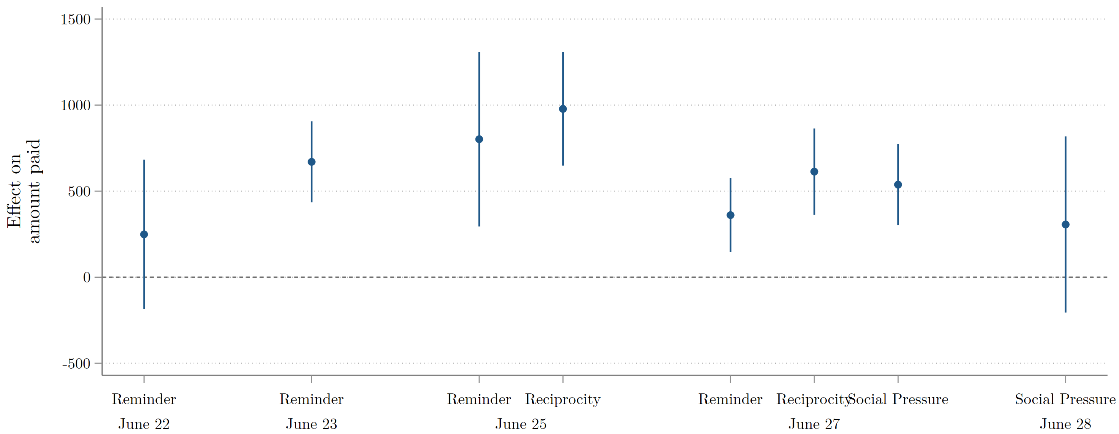
Note: Graph shows distribution of dates messages were sent out to the experimental sample.

Figure A2: Impact of receiving messages on different dates

(a) Outcome = any payment



(b) Outcome = amount paid

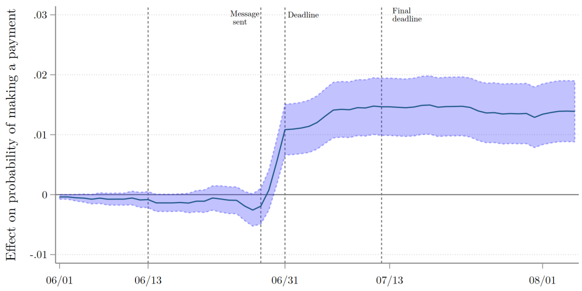


Notes: Taxpayers in the experimental sample received messages on different dates. To disentangle the timing of the messages from their individual effects, we run a (2SLS) regression of the impact of receiving one of each type of message with a dummy equal to one for one of the five days that most taxpayers received their message. Figure A2 shows the estimates from that regression. Note that on some days not every taxpayer received a message of every type. Taxpayers who received any messages before these dates are dropped from the specification. All dates are for 2018. 95% confidence intervals shown.

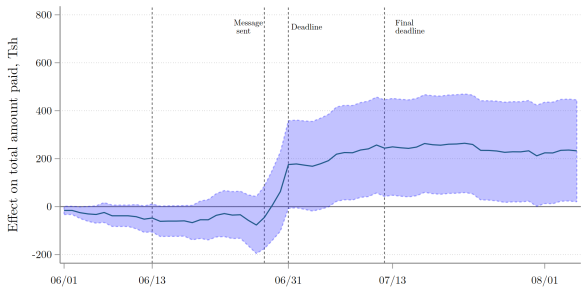
Figure A3: Time trajectory of outcomes for those that received a message on the same day

Outcome = any payment | **Outcome = amount paid**

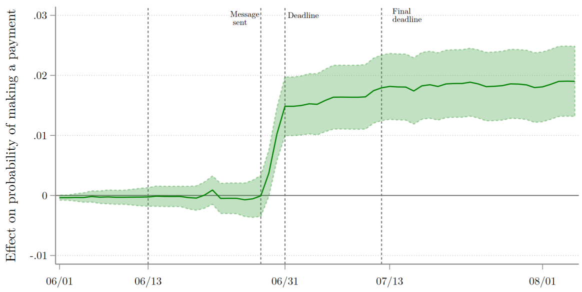
(a) Reminder



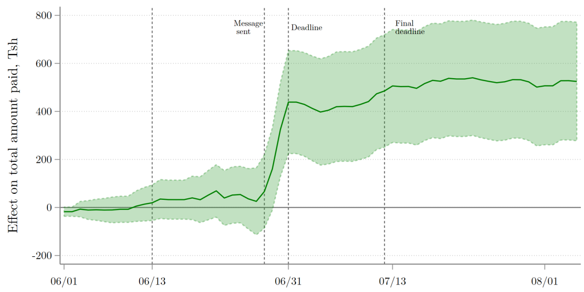
(d) Reminder



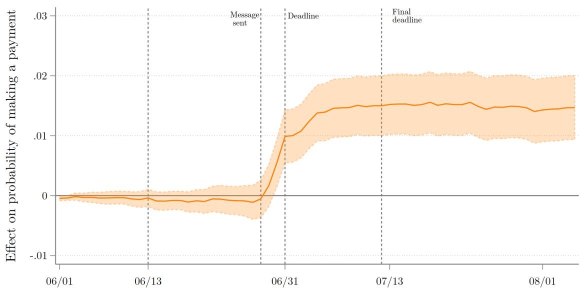
(b) Reciprocity



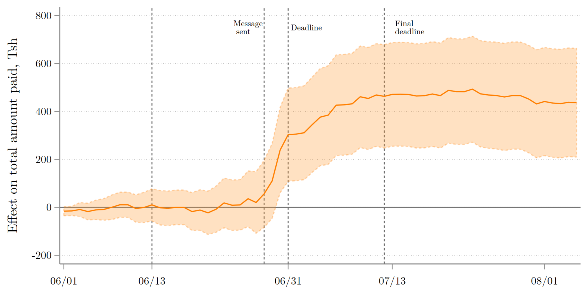
(e) Reciprocity



(c) Social Pressure

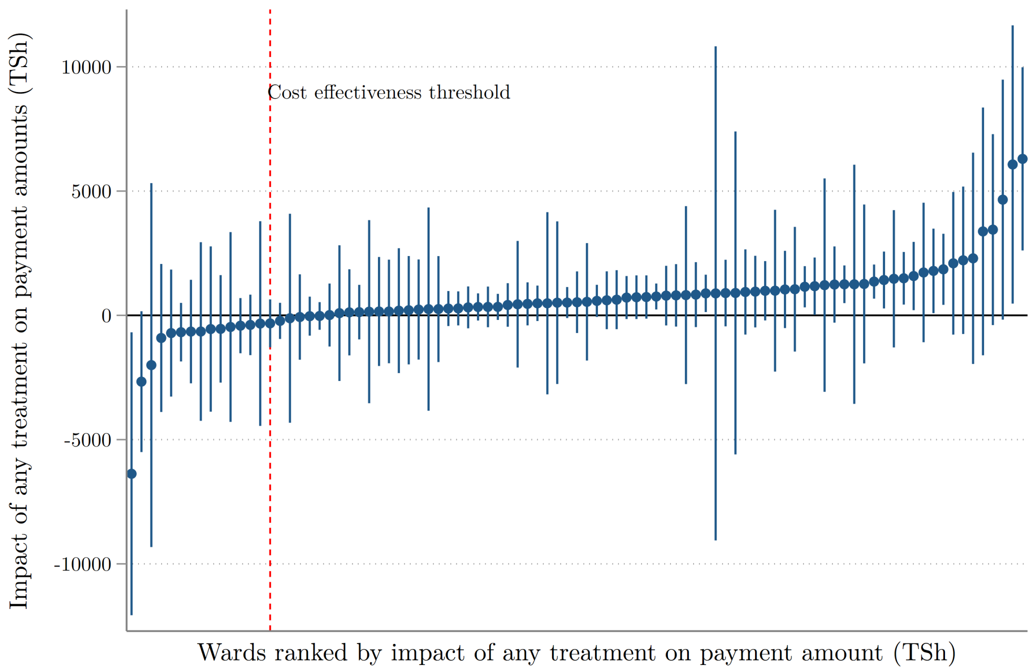


(f) Social Pressure



Notes: Each column displays the trajectory of 2SLS effect sizes for individuals who received their message on June 27th (the date that there is overlap between all three treatments) for both payment rates (the left column) and payment amounts (the right column). 95% confidence intervals shown.

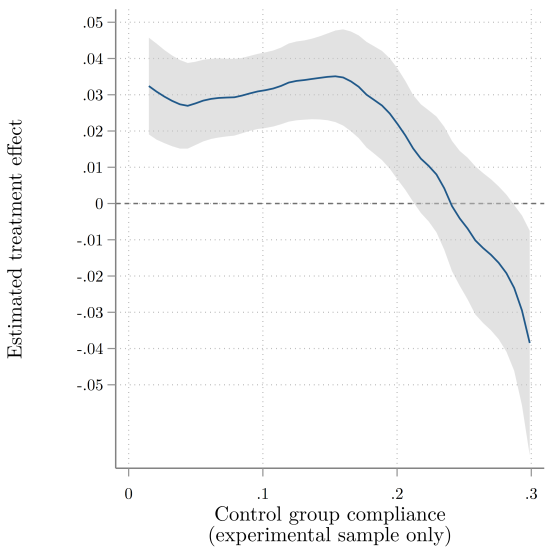
Figure A4: Impact of treatment on amount paid by ward



Note: Figure shows the effect size (estimated using 2SLS) of receiving any message, estimated separately for every administrative ward in Dar es Salaam. 95% confidence intervals shown.

Figure A5: Compliance in the control group or in the entire neighborhood as a predictor of the impact of the text message campaign

(a) Average compliance rate of the control group



(b) Predicted compliance rate of the entire ward without the intervention

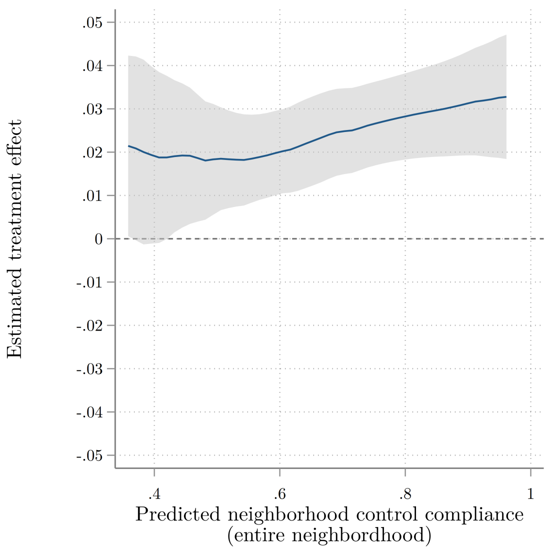
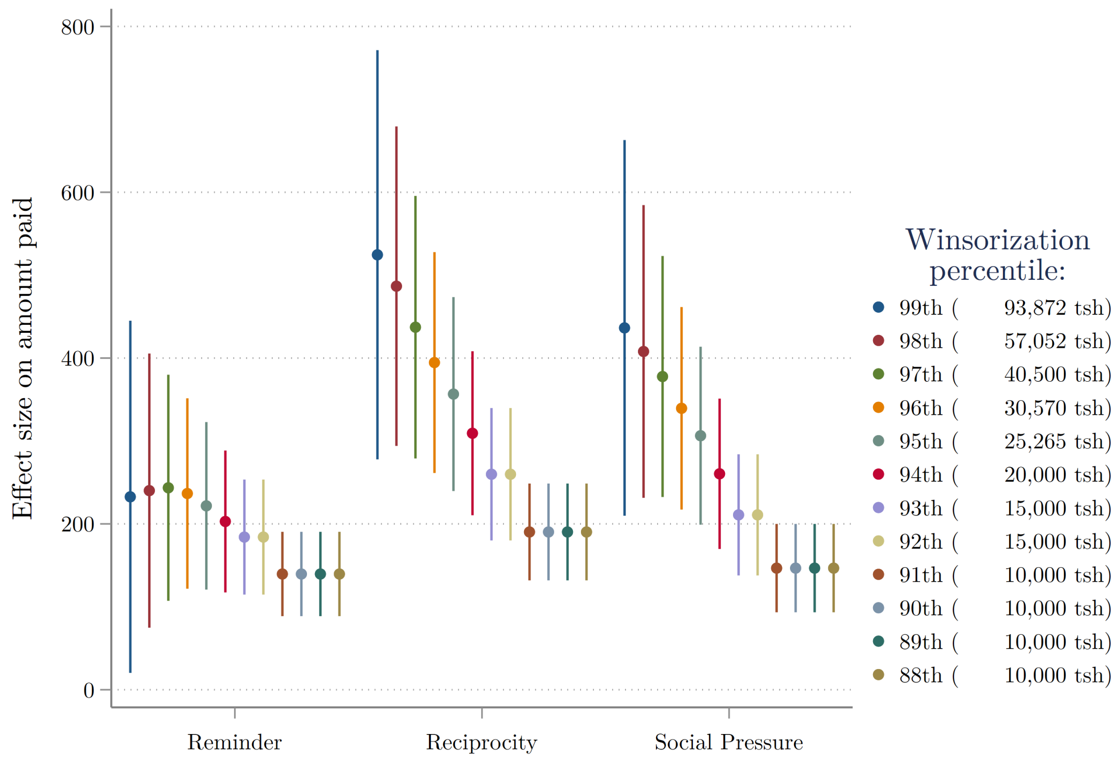


Figure A6: Effects on amount paid with different levels of winsorizing



Note: the above figure shows the estimated 2SLS impact of each treatment on payment amounts at different levels of winsorizing the outcome variable. The sample is restricted to the control group and indiduals who received a message on June 27th. 95% confidence intervals shown.

A1.2 Phone number spillovers and adjustments made to the final sample

The randomization was originally conducted over 237,699 taxpayer IDs, and the phone numbers associated with those IDs were used for the message treatments. But, as discussed above, the same phone number could be associated with multiple taxpayer IDs. This means that even if a taxpayer was only allocated into a single treatment arm (for example, the pure control), they might have inadvertently received a message intended for another treatment arm if they shared a phone number with a taxpayer in that arm.

Our estimates indicate that 39.62% of taxpayers share at least one other number with another taxpayer and, on average a taxpayer shares one other number with other taxpayers. Some taxpayers share several numbers: at the 99th percentile, a taxpayer shares a number with nine others.

Those that share more numbers are more likely, by chance, to receive multiple treatment assignments. The starkest example of this is for the control group: those with *any* shared numbers were 80% likely to receive an experimental message, where those without any shared numbers basically had no chance of receiving one.

However, a priori there is no reason to believe that this probability would vary across treatment arms. We can investigate this in two ways. First, Figure A7 shows the results of regressing the (inverse hyperbolic sine of the) number of shared numbers on each treatment arm (controlling for strata fixed effects), as well as a dummy equal to 1 if any numbers are shared. There appears to be no significant difference across treatment arms: those with more numbers do not appear to be more likely to be in the reminder treatment, versus the pure control, for example. Thus, we are able to restrict our analysis sample to taxpayers with unique phone numbers without undermining the randomization.

Figure A7: Differences in the number of shared phone numbers across treatment arms

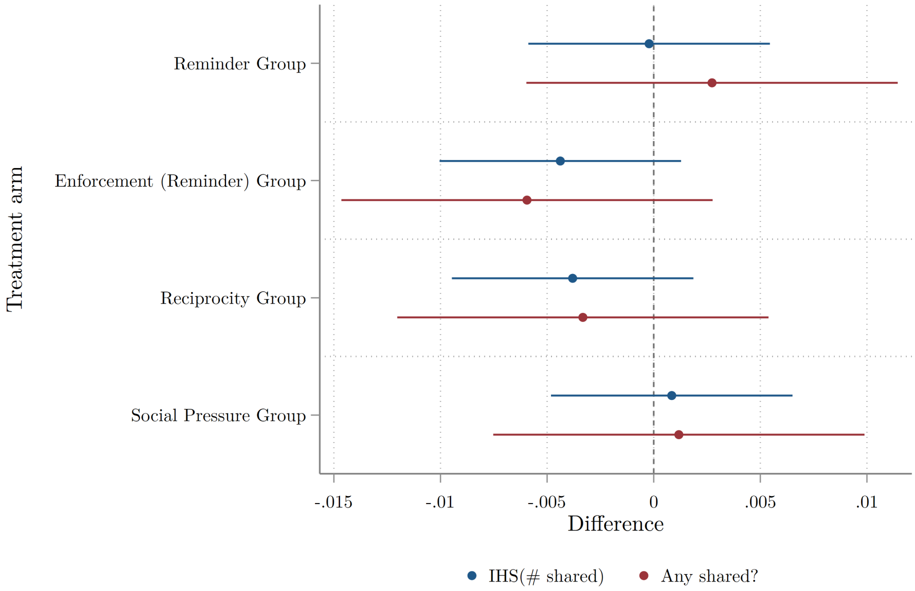
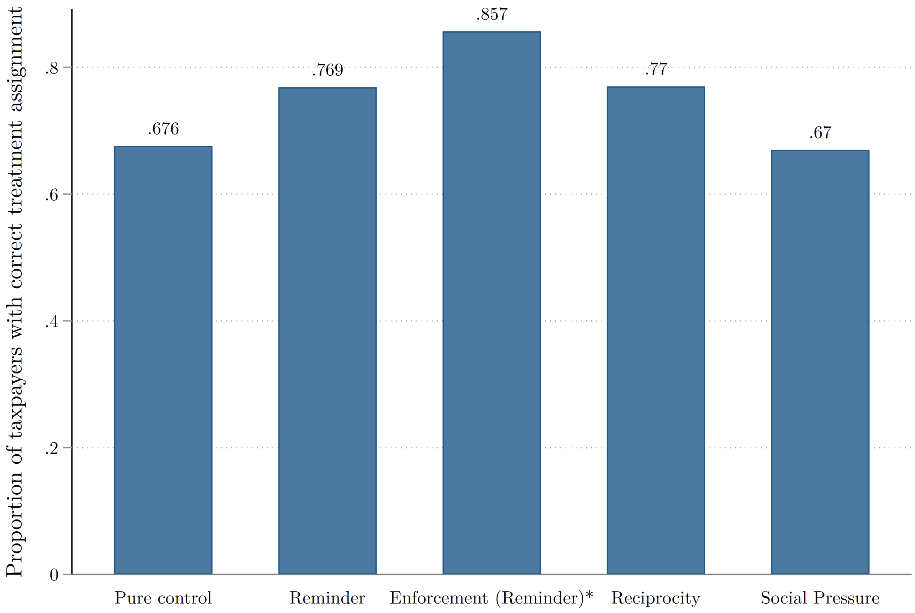
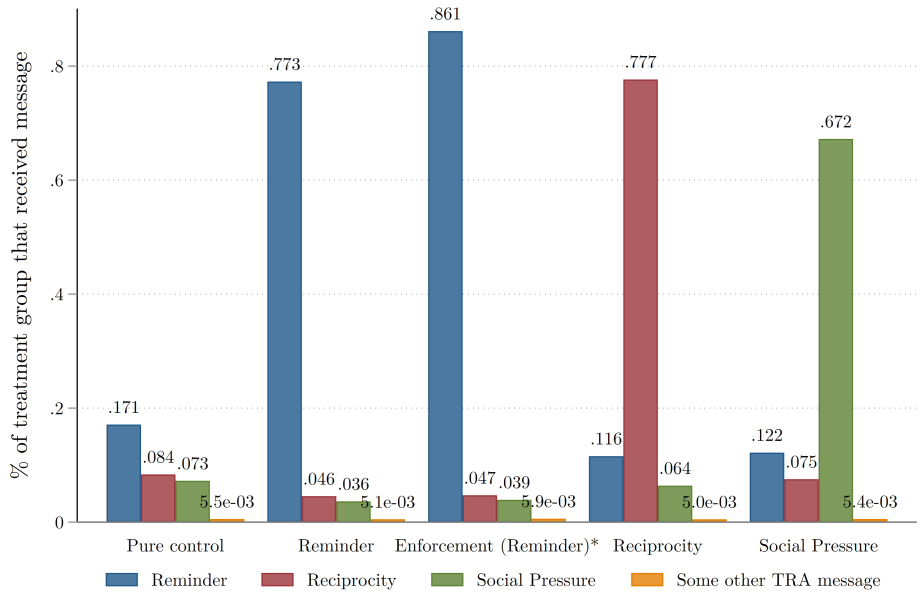


Figure A8: Proportion of taxpayers who received perfect treatment assignment



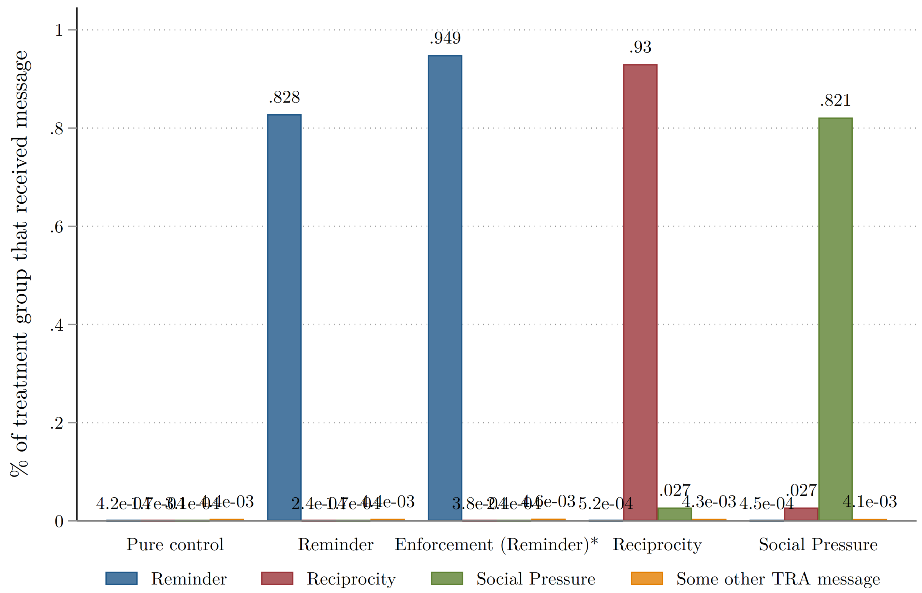
Note: Perfect assignment is defined as: receiving no experimental messages (pure control group), or receiving only the correct experiment message (all treatment groups) and no other. * While we include the enforcement treatment assignment group separately, because this group was inadvertently assigned to the reminder treatment, we judge perfect assignment here as only receiving a reminder message.

Figure A9: Distribution of received messages



Note: Perfect assignment is defined as: receiving no experimental messages (pure control group), or receiving only the correct experiment message (all treatment groups) and no other. * While we include the enforcement treatment assignment group separately, because this group was inadvertently assigned to the reminder treatment, we judge perfect assignment here as only receiving a reminder message.

Figure A10: Distribution of received messages when we restrict to taxpayers with a unique phone number



Note: Perfect assignment is defined as: receiving no experimental messages (pure control group), or receiving only the correct experiment message (all treatment groups) and no other. * While we include the enforcement treatment assignment group separately, because this group was inadvertently assigned to the reminder treatment, we judge perfect assignment here as only receiving a reminder message.